

Lecture Notes in Computer Science
Edited by G. Goos, J. Hartmanis and J. van Leeuwen

1923

Springer

Berlin

Heidelberg

New York

Barcelona

Hong Kong

London

Milan

Paris

Singapore

Tokyo

José Borbinha Thomas Baker (Eds.)

Research and Advanced Technology for Digital Libraries

4th European Conference, ECDL 2000
Lisbon, Portugal, September 18-20, 2000
Proceedings



Springer

Series Editors

Gerhard Goos, Karlsruhe University, Germany
Juris Hartmanis, Cornell University, NY, USA
Jan van Leeuwen, Utrecht University, The Netherlands

Volume Editors

José Borbinha
National Library of Portugal
Campo Grande, 83, 1749-081 Lisboa, Portugal
E-mail: Jose.Borbinha@bn.pt

Thomas Baker
GMD Library, Schloss Birlinghoven
53754 Sankt Augustin, Germany
E-mail: Thomas.Baker@gmd.de

Cataloging-in-Publication Data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Research and advanced technology for digital libraries : 4th European
conference ; proceedings / ECDL 2000, Lisbon, Portugal, September 18 -
20, 2000. José Borbinha ; Thomas Baker (ed.). - Berlin ; Heidelberg ;
New York ; Barcelona ; Hong Kong ; London ; Milan ; Paris ; Singapore ;
Tokyo : Springer, 2000
(Lecture notes in computer science ; Vol. 1923)
ISBN 3-540-41023-6

CR Subject Classification (1998): H.2, H.3, H.4.3, H.5, I.7, J.1, J.7

ISSN 0302-9743

ISBN 3-540-41023-6 Springer-Verlag Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution under the German Copyright Law.

Springer-Verlag Berlin Heidelberg New York
a member of BertelsmannSpringer Science+Business Media GmbH
© Springer-Verlag Berlin Heidelberg 2000
Printed in Germany

Typesetting: Camera-ready by author, data conversion by PTP-Berlin, Stefan Sossna
Printed on acid-free paper SPIN: 10722840 06/3142 5 4 3 2 1 0

Preface

ECDL2000, the Fourth European Conference on Research and Advanced Technology for Digital Libraries, is being held this year in Lisbon, Portugal, following previous events in Pisa (1997), Heraklion (1998), and Paris (1999). One major goal of the ECDL conference series has been to draw information professionals, stakeholders, and user communities from both the research world and from industry into a discussion of the alternative technologies, policies, and scenarios for global digital libraries. The success of previous conferences makes them a hard act to follow.

The field of digital libraries draws on a truly diverse set of scientific and technical disciplines. In the past three years, moreover, global cooperation on research and development has emerged as an urgent priority, particularly in the new European Framework Programme and in the Digital Library Initiative in the United States.

Because of this diversity, the field is perhaps still struggling for an identity. But this struggle for identity is itself a source of energy and creativity. Participants in this field feel themselves to be part of a special community, with special people. Each of us may claim expertise on a narrow issue, with specific projects, but the choices we make and the methods we use in local solutions can have unforeseen impacts within a growing universe of interconnected resources. This shared commitment to global interconnectedness gives the field a certain spirit, and this spirit in turn gives it a sense of coherence in the absence of a well-defined manifesto or established set of core themes. It also makes events such as ECDL2000 so very enjoyable!

This year we have tried to reflect this richness and diversity in the local organisation, in the Programme Committee, and in the final conference programme. The local organisation was handled by the National Library of Portugal; IST, the engineering institute of the Lisbon Technical University; and INESC, a research institution, with the additional support of DELOS, the European Network of Excellence for Digital Libraries. The Programme Committee includes members from 17 countries (eleven European Union countries plus Russia, USA, Australia, Japan, Morocco, and Brazil). The committee reflects a broad range of expertise from computer scientists and engineers to librarians, archivists, and managers.

Developing the conference programme has been a challenge. This year, 28 research papers address a cross-section of established topics in the field, while a new category of “short papers” provides an opportunity to explore emerging issues and re-evaluate the boundaries of our field. Soliciting short papers was an experiment, entailing some risk, but we felt this could help open the door to broader participation and perspectives. We are happy to see this interesting program complemented by some very high-quality panels, tutorials, keynote speakers, and workshops.

September 2000

José Borbinha
Thomas Baker

Organisation

ECDL2000 was organised in Lisbon, Portugal, by the National Library of Portugal, INESC (Instituto de Engenharia de Sistemas e Computadores), and IST (Instituto Superior Técnico – Universidade Técnica de Lisbon), and with the assistance of the DELOS Network of Excellence (supported by the Information Society Program of the European Commission).

General chair

José Tribolet, IST / INESC, Portugal

Programme and local chair

José Borbinha, BN / IST / INESC, Portugal

Workshops co-chairs

Thomas Baker, GMD, Germany
Alberto Silva, IST / INESC, Portugal

Tutorials co-chairs

Preben Hansen, SICS, Sweden
Eloy Rodrigues, Universidade do Minho, Portugal

Programme Committee

Philipe Agre, University of California, USA
Daniel E. Atkins, University of Michigan, USA
Joaquina Barrulas, INETI, Portugal
Rudolf Bayer, Technical University of Munich, Germany
Eurico Carrapatoso, University of Porto / INESC, Portugal
Rui Casteleiro, MIND S.A., Portugal
Panos Constantopoulos, FORTH, Greece
Ana M. R. Correia, UNL, Portugal
Bruce Croft, University of Massachusetts, USA
Murilo Bastos Cunha, University of Brasília, Brasil
Edward Fox, Virginia Tech, USA
Michael Freeston, University of Aberdeen, UK / University of California, USA
Norbert Fuhr, Dortmund University, Germany

VIII Organisation

Hachim Haddouti, debis Systemhaus Industry, Germany
Lynda Hardman, CWI, Netherlands
Cecília Henriques, AN / TT, National Archives, Portugal
Renato Iannella, IPR Systems Pty Ltd, Australia
Gareth Jones, University of Exeter, UK
Joaquim Jorge, IST / INESC, Portugal
Leonid Kalinichenko, Russian Academy of Sciences, Russia
Judith Klavans, Columbia University, USA
Carl Lagoze, Cornell University, USA
Alain Michard, INRIA, France
Marc Nanard, University of Montpellier, France
Erich J. Neuhold, GMD, Germany
Christos Nikolau, University of Crete / FORTH, Greece
Arlindo Oliveira, IST / INESC, Portugal
Mike Papazoglou, Tilburg University, Netherlands
Carol Peters, CNR, Italy
Alexander Plemnek, St. Petersburg State Technical University / Soros Foundation, Russia
Mogens Sandfær, DTV, Denmark
Leonel Santos, University of Minho, Portugal
João Sequeira, BAD / RTP, Portugal
Mário Silva, University of Lisbon, Portugal
Miguel Silva, IST / INESC, Portugal
Alan Smeaton, Dublin City University, Ireland
Ingeborg Solvberg, University of Trondheim, Norway
Shigeo Sugimoto, University of Library and Information Science, Japan
Constantino Thanos, CNR, Italy
Isabel Trancoso, IST / INESC, Portugal
Anne-Marie Vercoustre, INRIA, France
Paula Viana, ISEP / INESC, Portugal
Howard Wactlar, Carnegie Mellon University, USA
Stuart Weibel, OCLC, USA
Jian Yang, Tilburg University, The Netherlands

Reviewers

Philippe Agre, University of California, USA
Daniel E Atkins, University of Michigan, USA
Rudolf Bayer, Technical University of Munich, Germany
Eurico Carrapatoso, University of Porto / INESC, Portugal
Rui Casteleiro, MIND S.A., Portugal
Panos Constantopoulos, FORTH, Greece
Ana M. R. Correia, UNL, Portugal
Bruce Croft, University of Massachusetts, USA
Murilo Bastos Cunha, University of Brasília, Brasil
Nicola Fanizzi, University of Bari, Italy
Stefano Ferilli, University of Bari, Italy

Edward Fox, Virginia Tech, USA
 Nuno Freire, INESC / IST, Portugal
 Michael Freeston, University of Aberdeen, UK / University of California, USA
 Norbert Fuhr, Dortmund University, Germany
 Hachim Haddouti, debis Systemhaus Industry, Germany
 Cecília Henriques, AN / TT, National Archives, Portugal
 Renato Iannella, IPR Systems Pty Ltd, Australia
 Gareth Jones, University of Exeter, UK
 Joaquim Jorge, IST / INESC, Portugal
 Leonid Kalinichenko, Russian Academy of Sciences, Russia
 Carl Lagoze, Cornell University, USA
 Alain Michard, INRIA, France
 Marc Nanard, University of Montpellier, France
 Erich J. Neuhold, GMD, Germany
 Christos Nikolau, University of Crete / FORTH, Greece
 Arlindo Oliveira, IST / INESC, Portugal
 Mike Papazoglou, Tilburg University, Netherlands
 Carol Peters, CNR, Italy
 Alexander Plemnek, St. Petersburg State Technical University / Soros Foundation, Russia
 Mogens Sandfær, DTV, Denmark
 Leonel Santos, University of Minho, Portugal
 Giovanni Semeraro, University of Bari, Italy
 João Sequeira, BAD / RTP, Portugal
 Mário Silva, University of Lisbon, Portugal
 Miguel Silva, IST / INESC, Portugal
 Alan Smeaton, Dublin City University, Ireland
 Ingeborg Solvberg, University of Trondheim, Norway
 Shigeo Sugimoto, University of Library and Information Science, Japan
 Constantino Thanos, CNR, Italy
 Isabel Trancoso, IST / INESC, Portugal
 Anne-Marie Vercoustre, INRIA, France
 Paula Viana, ISEP / INESC, Portugal
 Howard Wactlar, Carnegie Mellon University, USA
 Jian Yang, Tilburg University, The Netherlands

Local Organising Committee

José Borbinha, BN / IST / INESC, Portugal
 Eulália Carvalho, BN, Portugal
 Fernando Cardoso, BN, Portugal
 João Leal, BN, Portugal
 João Maria Lopes, BN, Portugal
 Rita Alves Pereira, BN, Portugal
 Nuno Freire, INESC/IST, Portugal
 Hugo Amorim, IST, Portugal

Table of Contents

Research Papers

Optical Recognition

- Automatic Feature Extraction and Recognition for Digital Access of Books
of the Renaissance 1
*F. Muge, I. Granado, M. Mengucci, P. Pina, V. Ramos, N. Sirakov,
J. R. Caldas Pinto, A. Marcolino, Mário Ramalho, P. Vieira, and
A. Maia do Amaral*
- Content Based Indexing and Retrieval in a Digital Library of Arabic
Scripts and Calligraphy 14
Suliman Al-Hawamdeh and Gul N. Khan
- Ancient Music Recovery for Digital Libraries 24
*J. Caldas Pinto, P. Vieira, M. Ramalho, M. Mengucci, P. Pina, and
F. Muge*
- Probabilistic Automaton Model for Fuzzy English-Text Retrieval 35
Manabu Ohta, Atsuhiko Takasu, and Jun Adachi

Information Retrieval

- Associative and Spatial Relationships in Thesaurus-Based Retrieval 45
Harith Alani, Christopher Jones, and Douglas Tudhope
- Experiments on the Use of Feature Selection and Negative Evidence in
Automated Text Categorization 59
Luigi Galavotti, Fabrizio Sebastiani, and Maria Simi
- The Benefits of Displaying Additional Internal Document Information on
Textual Database Search Result Lists 69
Offer Drori
- Interactive-Time Similarity Search for Large Image Collections Using
Parallel VA-Files 83
Roger Weber, Klemens Böhm, and Hans-J. Schek

Metadata

- Dublin Core Metadata for Electronic Journals 93
Ann Apps and Ross MacIntyre
- An Event-Aware Model for Metadata Interoperability 103
Carl Lagoze, Jane Hunter, and Dan Brickley

XII Table of Contents

QUEST – Querying Specialized Collections on the Web	117
<i>Martin Heß, Christian Mönch, and Oswald Drobnik</i>	

Personal Data in a Large Digital Library	127
<i>José Manuel Barrueco Cruz, Markus J.R. Klink, and Thomas Krichel</i>	

Frameworks

Implementing a Reliable Digital Object Archive	128
<i>Brian Cooper, Arturo Crespo, and Hector Garcia-Molina</i>	

Policy-Carrying, Policy-Enforcing Digital Objects	144
<i>Sandra Payette and Carl Lagoze</i>	

INDIGO – An Approach to Infrastructures for Digital Libraries	158
<i>Christian Mönch</i>	

Scalable Digital Libraries Based on NCSTRL/Dienst	168
<i>Kurt Maly, Mohammad Zubair, Hesham Anan, Dun Tan, and Yunchuan Zhang</i>	

Multimedia

OMNIS/2: A Multimedia Meta System for Existing Digital Libraries	180
<i>Günther Specht and Michael G. Bauer</i>	

Modeling Archival Repositories for Digital Libraries	190
<i>Arturo Crespo and Hector Garcia-Molina</i>	

Implementation and Analysis of Several Keyframe-Based Browsing Interfaces to Digital Video	206
<i>Hyowon Lee, Alan F. Smeaton, Catherine Berrut, Noel Murphy, Seán Marlow, and Noel E. O'Connor</i>	

Functional and Intentional Limitations of Interactivity on Content Indexing Topics: Possible Uses of Automatic Classification and Contents Extraction Systems, in Order to Create Digital Libraries Databases	219
<i>Florent Pasquier</i>	

Users

Interaction Profiling in Digital Libraries through Learning Tools	229
<i>G. Semeraro, F. Esposito, N. Fanizzi, and S. Ferilli</i>	

DEBORA: Developing an Interface to Support Collaboration in a Digital Library	239
<i>David M. Nichols, Duncan Pemberton, Salah Dalhoumi, Omar Larouk, Claire Belisle, and Michael B. Twidale</i>	

Children as Design Partners and Testers for a Children's Digital Library . .	249
<i>Yin Leng Theng, Norliza Mohd Nasir, Harold Thimbleby,</i> <i>George Buchanan, Matt Jones, David Bainbridge, and Noel Cassidy</i>	
Evaluating a User-Model Based Personalisation Architecture for Digital News Services	259
<i>Alberto Díaz Esteban, Pablo Gervás Gómez-Navarro, and</i> <i>Antonio García Jiménez</i>	

Papers Complementing Invited Talks

Aging Links	269
<i>Claudia Niederée, Ulrike Steffens, Joachim W. Schmidt, and</i> <i>Florian Matthes</i>	
Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints	280
<i>Linda L. Hill</i>	
The Application of an Event-Aware Metadata Model to an Online Oral History Archive	291
<i>Jane Hunter and Darren James</i>	
From the Visual Book to the WEB Book: The Importance of Good Design	305
<i>M. Landoni, R. Wilson, and F. Gibb</i>	

Short-Papers

Multimedia

Topic Detection in Read Documents	315
<i>Rui Amaral and Isabel Trancoso</i>	
Map Segmentation by Colour Cube Genetic <i>K</i> -Mean Clustering	319
<i>Victorino Ramos and Fernando Muge</i>	
Spoken Query Processing for Information Access in Digital Libraries	324
<i>Fabio Crestani</i>	
A Metadata Model for Historical Documentary Films	328
<i>Giuseppe Amato, Donatella Castelli, and Serena Pisani</i>	
Image Description and Retrieval Using MPEG-7 Shape Descriptors	332
<i>Carla Zibreira and Fernando Pereira</i>	
A Large Scale Component-Based Multi-media Digital Library System - Development Experience and User Evaluation	336
<i>Hiroshi Mukaiyama</i>	

Users in Digital Libraries

Personalised Delivery of News Articles from Multiple Sources	340
<i>Gareth J. F. Jones, David J. Quested, and Katherine E. Thomson</i>	
Building a Digital Library of Web News	344
<i>Nuno Maria and Mário J. Silva</i>	
Automatically Detecting and Organizing Documents into Topic Hierarchies: A Neural-Network Based Approach to Bookshelf Creation and Arrangement	348
<i>Andreas Rauber, Michael Dittenbach, and Dieter Merkl</i>	
Daffodil: Distributed Agents for User-Friendly Access of Digital Libraries .	352
<i>Norbert Gövert, Norbert Fuhr, and Claus-Peter Klas</i>	
An Adaptive Systems Approach to the Implementation and Evaluation of Digital Library Recommendation Systems	356
<i>Johan Bollen and Luis M. Rocha</i>	
Are End-Users Satisfied by Using Digital Libraries?	360
<i>Mounir A. Khalil</i>	

Information Retrieval

CAP7: Searching and Browsing in Distributed Document Collections	364
<i>Norbert Fuhr, Kai Großjohann, and Stefan Kokkelink</i>	
Representing Context-Dependent Information Using <i>Multidimensional</i> <i>XML</i>	368
<i>Yannis Stavrakas, Manolis Gergatsoulis, and Theodoros Mitakos</i>	
AQUA (Advanced Query User Interface Architecture)	372
<i>László Kovács, András Micsik, Balázs Pataki, and István Zsámboki</i>	
Fusion of Overlapped Result Sets	376
<i>Joaquim Macedo, António Costa, and Vasco Freitas</i>	
ActiveXML: Compound Documents for Integration of Heterogeneous Data Sources	380
<i>João P. Campos and Mário J. Silva</i>	
newsWORKS©, the Complete Solution for Digital Press Clippings and Press Reviews: Capture of Information in an Intelligent Way	385
<i>Begoña Aguilera Caballero and Richard Lehner</i>	

Internet Cataloguing

Effects of Cognitive and Problem Solving Style on Internet Search Tool . .	389
<i>Tek Yong Lim and Enya Kong Tang</i>	

Follow the Fox to Renardus: An Academic Subject Gateway Service for Europe	395
<i>Lesly Huxley</i>	
CORC: Helping Libraries Take a Leading Role in the Digital Age	399
<i>Kay Covert</i>	
Automatic Web Rating: Filtering Obscene Content on the Web	403
<i>K. V. Chandrinos, Ion Androustopoulos, G. Paliouras, and C. D. Spyropoulos</i>	
The Bibliographic Management of Web Documents in Digital and Hybrid Libraries	407
<i>Wallace C. Koehler, Jr.</i>	

Technical Collections

The Economic Impact of an Electronic Journal Collection on an Academic Library	413
<i>Carol Hansen Montgomery and John A. Bielec</i>	
A Comparative Transaction Log Analysis of Two Computing Collections ..	418
<i>Malika Mahoui and Sally Jo Cunningham</i>	
ERAM - Digitisation of Classical Mathematical Publications	424
<i>Hans Becker and Bernd Wegner</i>	
The Electronic Library in EMIS - European Mathematical Information Service	428
<i>Bernd Wegner</i>	
Model for an Electronic Access to the Algerian Scientific Literature: Short Description	432
<i>Bakelli Yahia</i>	

Cases 1

A Digital Library of Native American Images	437
<i>Elaine Peterson</i>	
China Digital Library Initiative and Development	441
<i>Michael B. Huang and Guohui Li</i>	
Appropriation of Legal Information: Evaluation of Data Bases for Researchers	445
<i>Céline Hembise</i>	

XVI Table of Contents

Publishing 30 Years of the Legislation of Brazil's São Paulo State in CD-ROM and Internet	449
<i>Paulo Leme, Dilson da Costa, Ricardo Baccarelli, Maurício Barbosa, Andréa Bolanho, Ana Reis, Rose Bicudo, Eduardo Barbosa, Márcio Nunes, Innocêncio Pereira Filho, Guilherme Plonski, and Sérgio Kobayashi</i>	

Electronic Dissemination of Statistical Information at Local Level: A Cooperative Project between a University Library and Other Public Institutions	452
<i>Eugenio Pelizzari</i>	

Cases 2

Building Archaeological Photograph Library	456
<i>Rei S. Atarashi, Masakazu Imai, Hideki Sunahara, Kunihiro Chihara, and Tadashi Katata</i>	

EULER - A DC-Based Integrated Access to Library Catalogues and Other Mathematics Information in the Web	461
<i>Bernd Wegner</i>	

Decomate: Unified Access to Globally Distributed Libraries	467
<i>Thomas Place and Jeroen Hoppenbrouwers</i>	

MADILIS, the Microsoft Access-Based Digital Library System	471
<i>Scott Herrington and Philip Konomos</i>	

Leveraging Electronic Content: Electronic Linking Initiatives at Arizona State University	475
<i>Dennis Brunning</i>	

Cases 3

Asian Film Connection: Developing a Scholarly Multilingual Digital Library - A Case Study	481
<i>Marianne Afifi</i>	

Conceptual Model of Children's Electronic Textbook	485
<i>Norshuhada Shiratuddin and Monica Landoni</i>	

An Information Food Chain for Advanced Applications on the WWW	490
<i>Stefan Decker, Jan Jannink, Sergey Melnik, Prasenjit Mitra, Steffen Staab, Rudi Studer, and Gio Wiederhold</i>	

An Architecture for a Multi Criteria Exploration of a Documents Set	494
<i>Patricia Dzeakou and Jean-Claude Derniame</i>	

An Open Digital Library Ordering System	498
<i>Sarantos Kapidakis and Kostas Zorbadelos</i>	

Special Workshop

Special NKOS Workshop on Networked Knowledge Organization Systems . 502
Martin Doerr, Traugott Koch, Douglas Tudhope, and Repke de Vries

Implementing Electronic Journals in the Library and Making them
 Available to the End-User: An Integrated Approach 506
*Gerrit Alewaeters, Serge Gilen, Paul Nieuwenhuysen, Stefaan Renard,
 and Marc Verpoorten*

Author Index 511

Automatic Feature Extraction and Recognition for Digital Access of Books of the Renaissance

F. Muge¹, I. Granado¹, M. Mengucci¹, P. Pina¹, V. Ramos¹, N. Sirakov¹, J.R. Caldas
Pinto², A. Marcolino², Mário Ramalho², P. Vieira², and A. Maia do Amaral³

¹CVRM / Centro de Geo-Sistemas, Instituto Superior Técnico,
Av. Rovisco Pais, 1049-001 Lisboa

²IDMEC, Instituto Superior Técnico, Av. Rovisco Pais, 1049-001 Lisboa

³Biblioteca Geral da Universidade de Coimbra, Largo da Porta, 3000 Coimbra

Abstract. Antique printed books constitute a heritage that should be preserved and used. With novel digitising techniques is now possible to have these books stored in digital format and accessible to a wider public. However it remains the problem of how to use them. DEBORA (Digital accEss to BOoks of the RenAissance) is a European project that aims to develop a system to interact with these books through world-wide networks. The main issue is to build a database accessible through client computers. That will require to build accompanying metadata that should characterise different components of the books as illuminated letters, banners, figures and key words in order to simplify and speed up the remote access. To solve these problems, digital image analysis algorithms regarding filtering, segmentation, separation of text from non-text, lines and word segmentation and word recognition were developed. Some novel ideas are presented and illustrated through examples.

Introduction

Among invaluable library documents are ancient books, which, beyond their rarity and value, represent the 'state of knowledge'. They constitute a heritage that should be preserved and used. These two considerations represent conflicting aspects that should be overcome. How to use antique books, without damaging and degrading them? One response is to store them on a suitable media format and to allow their use over a suitable system. Currently there is no doubt that the suitable media is the digital media support and the system to access it through computer networks.

The fact that books must be handled with care to avoid damage does not allow that pages are flattened on the digitisation process, and thus geometrical distortions may appear. In most books, pages may be degraded due to humidity or other natural causes. This implies a pre-processing phase that includes geometric correction and filtering. The book is then ready to be accessed. However it is important to build an ergonomic interface that permits to an user to interact with books, allowing their efficient consultation, for example about the presence or absence of certain specific graphic characteristics or even given keywords. For that, other steps have to be carried out: one is page decomposition upon text, figures, illuminated letters and banners, while the other step is to search and to compare graphemes.

Gutenberg invented the printing press in the middle 1400's. However it remained some centuries until graphical types were normalised and automatically done, allowing type regularity. Effectively XVI century books are irregular in format and graphic quality that make them unique exemplars. This address particular problems to recognition, which, cannot be supported by a dictionary or grammatical database.

Different work has been done in regard to this area of document image analysis. Segmentation is only a step highly dependent on the objectives and requirements proposed [4,7,8,16]. Thus, it may appear that one could pick up an assembled component of software, use it like a black box, and get the job done. That would be nice, but unfortunately there are no such 'building blocks' available.

In this paper is presented a methodology that integrates several digital image analysis techniques to automatically extract and recognise features from books of the Renaissance. After image acquisition procedures, it consists on five main phases: (1) page segmentation, (2) separation of text from non-text, (3) line detection, (4) word detection and (5) word recognition.

1 Page Segmentation

The image acquisition task, i.e., the creation of digital images of books' in order to become understandable by computers, was performed using the DigiBook scanner from Xerox. The images were acquired in grey level with some pre-processing operations, namely geometrical corrections and filtering to attenuate noise.

Using the grey level images (one digital image refers to one complete page of the book), a mathematical morphology based methodology for page segmentation was developed. Readers are advised to consult Serra [13] and Soille [14] for the definitions. The concept guiding this methodology to segment the images refers to a topographic analogy. The signs (text and figures) to detect are like "valleys" or "furrows" (darker regions) carved on a basically plane background (lighter regions). To detect them, they are filled and the original image is subtracted to that one. To locate the interesting regions of each image, in order to build an appropriate mask, the morphological gradient information was exploited. This would leave out of the process all the objects (dirt spots, etc) with "soft slope" contours [2].

In the following, basic directional morphology operators to suite the text's shape and the figures were used [1]. The mask is therefore obtained through directional dilations of the gradient, in the horizontal and vertical directions. The intention was to make the minimum number of dilations needed to cover the "furrow" of any sign (text or figure) to be detected, and not more, otherwise the "valleys" of surrounding noise could be covered. Moreover, it was assumed that the main directions of the text characters were 0° and 90° , which is close enough to the reality to get effective results. This method gives an output image with a dark background and the extracted objects in grey level. It is, basically, a sequence in 3 steps: (1) dilating the original grey level image, (2) reconstruction of the original image with the image created in step one, which is the "covering" or "filling" operation, (3) subtraction of the original image to the reconstructed one. It was considered convenient to repeat the operation once more, as letters or figures lightly marked were badly segmented with the first gradient mask. The second "covering" is somehow complementary to the first one. The reason is the following: after the dilated gradient is processed, an image created by the

multiplication of the binary mask obtained by threshold is obtained (called the dilated gradient obtained by the maximum value of the gradient, meaningfully called “plateau”). Then, the gradient is subtracted to this “plateau”. The result is an image with minima located where the dilated gradient presented maxima and vice-versa. Next, this “complementary dilated gradient” is raised until its maximum meets the value of the first gradient’s maximum. This allows the second reconstruction to cover and fill the “valleys” of light marks, this being a problem especially in areas with a very white, or “high” in topographic terms, background. The algorithm can be consulted in detail in Mengucci and Granado [10]. The application of this algorithm is presented in figure 1. This example on a page of the first edition of *Os Lusíadas* (1572) by Luís de Camões, illustrates the automatic creation of a binary image (final step of the algorithm) starting from a grey level one.



Fig. 1. Page Segmentation applied to *Os Lusíadas* Book : (a) initial and (b) final images

2 Separating Text from Non-text

Based on the segmented pages, the objective in this phase is to separate non-text (figures, labels, illuminated letters) from text characters. Using the binary images a mathematical morphology based algorithm was developed. Fundamentally, uses the main directions in which the image’s elements evolve, i.e., assuming that the text has an orientation of 0° (length) and 90° (height) in the majority of lines composing the characters.

Most of the figures are composed of several small lines, whose main directions do not include the ones concerning text characters. So, the figures were obtained using the directions of 30°, 60°, 120° and 150°. All the images were converted to a hexagonal grid, in order to have smoother shapes in the results (due to its isotropic feature). The skew of the images was not taken into account because viewing our image set we noticed a regular behaviour, with no perceptible rotation [11].

Concerning the algorithm's implementation, it was used a combination of mathematical morphology operations/primitives to obtain the expected separation or segmentation. It starts with the figure's extraction followed by the separation of the text.

The sequence of the algorithm is described in the following (steps 1-9 respect figure segmentation, while steps 10-15 respect text segmentation):

1. After the conversion to a hexagonal grid, directional closings at 30°, 60°, 120° and 150°, with structuring element size of adequate size, to connect figure elements. The size chosen must avoid the connection between adjacent characters.
2. Directional openings, at 0° and 90°, eliminating most of the text characters. At this phase, according to the parameter's values all the figures are separated from the rest.
3. Reconstruction of the image resulting from step 1, using as a marker the image resulting from the step 2.
4. Reconstruction of the initial image using as markers the output of step 3.
5. Directional closings applied to the output of step 4 with a structuring element of size 3.
6. Hole fill operation to the output of step 5, closing all the figures elements.
7. Opening applied to the output image of step 6, excluding all the remaining text elements.
8. Reconstruction of the image resultant from step 6, using as marker the output of step 7.
9. Reconstruction of the initial image, using as markers the output of step 8, obtaining the figures.
10. Logical asymmetrical subtraction, between the image resultant from step 9 and the initial one. This way it is obtained the respective text, as well as some small structures that should be cleaned.
11. Suppression of the elements touching the image's border.
12. Directional openings over the output of step 11, at 30°, 60°, 120° and 150° and directions 0° and 90°, to eliminate the small particles.
13. Directional closing to the output of step 12, at 0°, 30°, 60°, 90°, 120° and 150° directions, in order to connect the text characters.
14. Reconstruction of output of step 11 using as markers the output of step 13.
15. Initial image reconstruction using as markers the output of step 14, obtaining the text cleaned from noisy particles.

Although the algorithm is designed for general application, it is necessary to adapt its parameter values to each type of book, due to its typographical and printing features as well as its preservation conditions.

Two examples illustrating the methodology to separate text from non-text are shown in figures 2 and 3. In both cases the non-text and text elements are clearly extracted, even if, in the *Os Lusíadas* case (figure 2a), a letter "Q" is classified as part of the non-text set (figure 2b). Errors like this one are consequence of a deficient binarisation, which is usually a very critical point [5]. In the other example (figure 3a), it can be seen that the evident grey level variations of the background, were not an obstacle to a good binarisation using the developed algorithm.

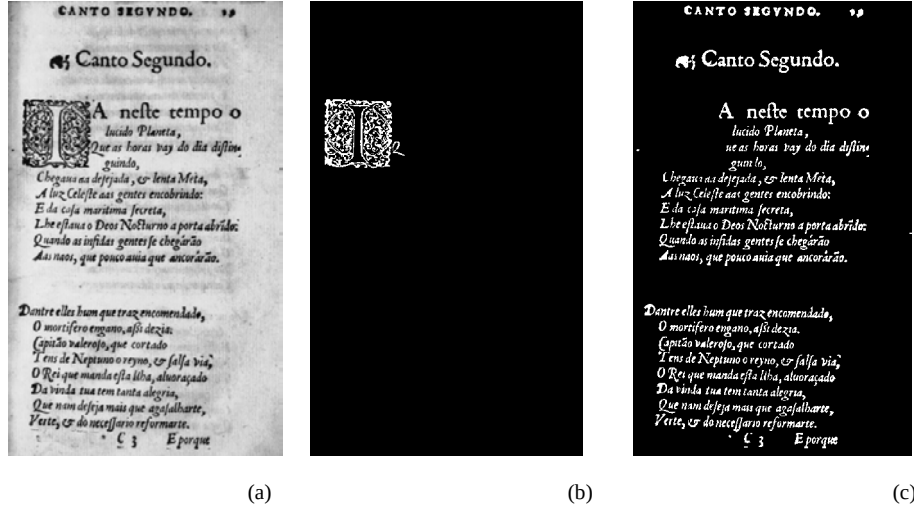


Fig. 2. Example of separation of text from non-text in a page of Os Lusíadas book. (a) initial, (b) non-text and (c) text images

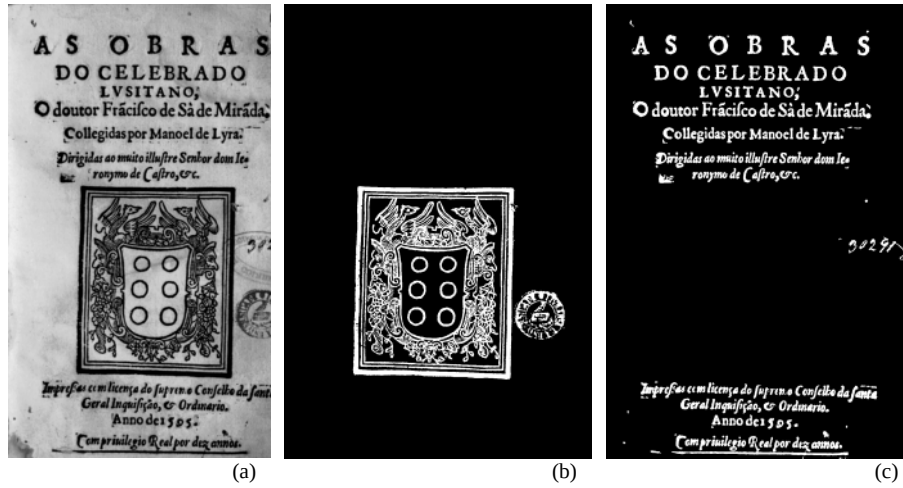


Fig. 3. Example of separation of text from non-text in the front page of Obras do Celebrado Lusitano book: (a) initial, (b) non-text and (c) text images

3 Line Detection

The process of line detection is performed upon the binary images resulting from the previous algorithm, i.e., on the text image only. Through this procedure, the semantic level of the document increases, once having a simple binary image in the input of the algorithm, one gets the knowledge about the lines of text that exist. Thus, at the end of

this step, further analysis of the document will not be performed only on raw input, but is also carried out on a line by line basis.

To achieve the task of line segmentation, a simple and fast algorithm giving the expected results was developed. It starts by computing the projection profile across horizontal lines of the image. Based on this projection profile, a special kind of operation that leads to the solution is performed, involving the computation of a threshold that will serve to extract the lines of text. The used binary images present noise and degradation due to a non-optimal binarisation procedure (still under improvement) which implies further difficulties. The noise in the image, which might cause unexpected behaviour, is treated with the aid of a ‘democratic solution’.

When it is intended to extract the lines, the data that characterise a line must be previously defined. It was decided that the following four features are the necessary ones: the baseline, the x-line, the bottom line and the top line (figure 4). These features are the ones required due to nature of text being analysed upon.

Usually those lines follow approximately the following criterion:

$$(\text{top\_line} - \text{x\_line}) = (\text{x\_line} - \text{baseline}) = (\text{baseline} - \text{bottom\_line}) = k. \quad (1)$$

The early binding of this type of restriction has allowed us to speed up the process, at the cost of restringing the segmentation process to similar typographic structures to the one that are envisaged. To process another type of typographic structures, this approach may face some difficulties, that can be easily corrected, but at computing time expenses.

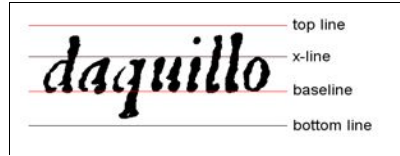


Fig. 4. Features to analyse

Projection profile is a simple operation: for each line in the image the number of black pixels that appear in that line are computed. Thus, at the end of applying this operation to all the lines, the vector of values that is obtained constitutes a histogram. Figure 5 presents this projection profile, superimposed to the original image.

Lines are extracted in the following by segmentation within appropriate thresholds the above histograms. However, test samples do not present homogeneous characteristics in resolution or scale. This does not allow the threshold values to be fixed 'a priori', being needed, in order to simplify the process, to fix an automatic criterion. Looking carefully at figure 5 one can see that the projection profile almost intuitively gives the lines of text that are formed by the image. Imagining a simple vertical line that crosses the entire image at the upper side of the projection profile and that by walking through that hypothetical line from the top until the bottom, two distinct areas may be created several times. One area will be “empty” in the histogram, and the other area will be “full”. The transitions between the “empty and full areas” allow detecting the baseline and the x-line of each line of text.

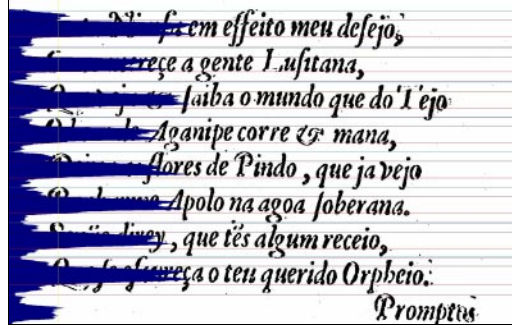


Fig. 5. Horizontal Projection Profile

To compute the location of this line, a statistical measure called the “mean value of pixels per line” is introduced. It is computed through the following expression:

$$\overline{mvpl} = \frac{\sum_{i=1}^{Height} x^i}{Height} \quad (2)$$

In the example of figure 5, a total of 8 lines were extracted. This approach still leaves some cases unattended. Unexpected shorter lines, as the last one shown in figure 5 are not detected. However, this error can be easily corrected if the same algorithm is reapplied to the portions of the image that did not contribute with any line. This task is facilitated if the overall page is already segmented in blocks of text. A second kind of problem arises when walking through the projection profile, along the mean value. A single line may be cut in two or more 'lines'. This problem is corrected through a voting procedure. Each line votes with the corresponding height to choose a representative. The majority elects the representative lines and those that significantly differ (for instance, >50%) from the representative are discarded.

4 Word Detection

The objective now is to separate the different words that exist in each line of the page. The process of word detection or segmentation consists on processing a binary image containing only text characters and the respective detected lines. It must stressed out, that also in this case the semantic level of the document is increased. After this task, one not only has the raw images and the lines of text that are part of the image, but also has knowledge about the words that are in the lines of text.

The procedure starts by computing the projection profile of the image, iterating across vertical lines. With this projection profile, the histogram of the space length is build, where two distinct areas are clearly present. The space between characters corresponds to the smaller bin values, while the spaces between words correspond to higher bin values. Thus, it is necessary to find out an objective criterion for automatic thresholding.

The projection profile is almost identical to the previous one: for each vertical line in the image, the number of black pixels that appear in that line is computed. At the end, the vector of values gives a histogram. In figure 6, a representation of this projection profile and its superimposition to the original image are presented.

It is important to note that the interest now is to represent a word by the involved bounding box. Therefore, a rectangle or two extremities of a main diagonal can represent the bounding box.



Fig. 6. Word detection: (a) input line and (b) superimposed projection profile

To solve the problem of word segmentation a novel approach to split words apart is introduced. The approach relies on a length assumption that can be seen as rules to achieve a solution. It is assumed that two characters are separated by a very small amount of pixels and that two words are separated by a significant amount of pixels. With the projection profile it is possible to construct a histogram of the lengths of white spaces that are in the line. An example can be seen in figure 7. Analysing this histogram, it may be considered that its left side corresponds to the spaces between the characters (small widths), and that its right side corresponds to the spaces between the words (bigger widths). The final result can be seen in figure 8, where 6 words were extracted.

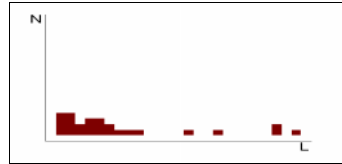


Fig. 7. Word length histogram

It is assumed that there are spaces between characters and that they are bigger than spaces between words. As a consequence of a poor binarisation or text formatting sparsely, some words can be split apart. On the other hand words may be aggregated, as word spacing may present irregularities in this type of documents. In this case, the algorithm could present a serious failure rate if the relative spacing assumption is not verified. However, because this operation is carried out line by line it is, in general, possible to cope with space variations in these ancient printed documents.

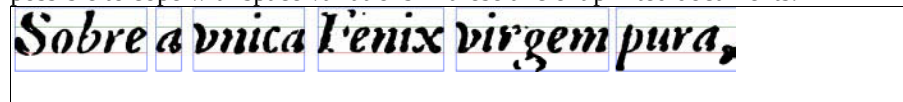


Fig. 8. Word detection

5 Word Recognition

Several solutions have been proposed with the specific job of doing word recognition, which is related in some ways with character recognition. However, the difficulty of this type of task is enormous, and increases directly with the degree of degradation of the scanned text and figures. Ideally, it is expected to convert the text to an ascii format, a typical feature of an OCR procedure. But, what in reality happens is that the analysed texts contain symbols that are not used anymore. The respective conversion in this way is not possible or not desired, once the recognition rates are very low.

One of the techniques to carry out the recognition consists of extracting a set of features from the pictures, and then to use that information to obtain a good classification. Several features have been proposed in document analysis literature (strokes, contour analysis, etc.). However, the quality of the results quickly deteriorates as the degradation of the picture increases.

A review of some methodologies has led us to test them, as presented in the following to our case studies. We can somehow anticipate that only a combination of features will tend to better solve the word recognition problem.

The information about the bounding boxes that limit the words is the departure point. The task involves selecting a picture of a word and presenting all the words that look alike.

5.1 Word Length Threshold

This is the simplest approach, being obvious that a matching candidate word should approximately have the same length as the word to match. Using images scanned at 300 dpi it was found out that a tolerance of 10 pixels will allow considering all the possible candidates.

Correlation between Words

Another feature that should be considered is the spatial correlation between words, if there are no severe problems of distortions or rotations between words (those have been eliminated in the pre-processing stage). The following measure (SSD) between images I_1 and I_2 was used:

$$SSD = \frac{\sum (I_1 - I_2)^2}{\sqrt{\sum I_1^2 \sum I_2^2}} \quad (3)$$

Results have shown that correlation is a good measure and should be heavily weighted. Another advantage of this method is fast computation time, which is of great importance when dealing with an entire book of hundreds of pages. Results for some words are shown in figure 9. In certain cases the discrimination almost suggests that no other measures are needed. However, as it can be seen in the same figure, the existence of false positives, is a warning to be cautious and therefore look for another descriptors.

5.2 Character–Shape Code (CSC)

Another method to help in discriminating candidates is the so-called character shape code [15], where each character image is divided in three zones (figure 10) and classified accordingly.

Keyword :	autres	sciences	toutes	bien	vous	Roy
Matches :						
1	autres	sciences	toutes	bien	vous	Roy
2	autres	sciences	toutes	bien	vous	long
3	autres	tenebres	toutes	pieu	vous	mon
4	autres	quelque	mon	meu	vous	Del-
5	autres	discours	forces	bien	vous	bloit
6	autres	Leuant	quels	plan	vous	faire
7	autres	speciale	renom	bien	vous	leurs
8	apres	fuiuant	nuyet	met	vous	long
9	auons	fuiuoien	notre	bien	vous	tant
10	autres	aucune-	viuant	plus	nom	mon

Fig. 9. Matching results using correlation results are then ordered by increasing order of similarity

Characters that have pixels between the top line and the x-line are considered ascenders. Characters that have pixels between the baseline and the bottom line are considered descenders (like the second character in figure 10). Lastly, characters that have all the pixels between the x-line and the baseline are neither ascenders nor descenders. Variations of this codification are described in detail in the paper of Splitz [15].

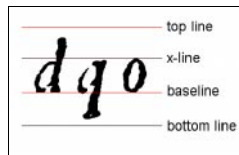


Fig. 10. Features in a character image

A signature based on ascenders and on descenders can this way be extracted and possibly used to select or discard words. Unfortunately, broken and touching characters occur in a non-neglected frequency. The codification in ascenders and descenders in these cases may not be applied character by character. This method was

then modified to extract the information about ascenders and descenders at fixed points in a word's image, bypassing the character segmentation phase. Each word can be divided in sectors with a pre-specified width (approximately character width), as illustrated in figure 11. For this image, the codified string would be "Aaxxxxx". An 'A' designates a sector with an ascender. A 'x' designates a sector that is neither an ascender nor a descender. A 'D' designates a sector that is a descender.

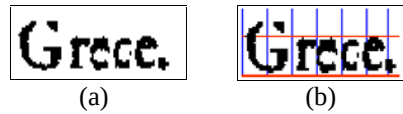


Fig. 11. Shape code string evaluation

A simpler measure consists on computing the number of mismatches between the shape codes under comparison. As already stated out, the words that have a poor binarisation normally cause problems in the extraction of the signatures. With incorrect signatures taken from the words, this method is prone to errors that were not desired. It may be concluded that this method may help to detect some intruders in the list obtained by correlation.

5.3 Ulam's Distance (UD)

Recently, Bhat [3] has presented an evolutionary measure for image matching. It is based on the Ulam's distance - a well know ordinal measure from molecular biology, based on an evolutionary distance metric that is used for comparing real DNA strings. Given two strings, the Ulam's distance is the smallest number of mutations, insertions, and deletions that can be made within both strings such that the resulting substrings are identical. Then, Bhat [3] reinterprets the Ulam's distance with respect to permutations that represent windows intensities expressed on an ordinal scale. His motivation to use this measure is twofold: to give a robust measure of correlation between windows and to help in identifying pixels that contribute to the agreement (or disagreement) between the windows. This approach, even if robust, has some drawbacks. However, they are concerned with grey level nature of the image. [9,12] it is presented an adaptation of this algorithm to binary images, with good results.

5.4 Character Segmentation (CS)

The aim here is not to perform an OCR based on character segmentation. However, all the techniques developed to word segmentation can be applied to segment the characters in a given word. The number of obtained characters (even if incorrect due to typographical specificities) may be another feature that can contribute to the ultimate decision algorithm.

5.5 Results

The algorithms reviewed in this paper will be tested in order to check their possible contribution to a decision rule based on a fusion of a set of descriptors. A small example is presented in table 1: in the first column, words are aligned according with the correlation results, while in the remaining ones is presented for each case a measure of distance from the corresponding word to the respective input keyword. It is expected that this measure give us some insight about how these features may help to correct results obtained by the correlation technique.

Keyword :		CSC	UD	CS
Matches :		0	24.8	0
1	delcription	0	28.0	0
2	delcription	4	30.6	2
3	demeurera	6	28.6	1
4	s'efforçants			

Table 1. Results for the word „description“

6 Conclusions

This paper dealt with document extraction and recognition techniques applied to XVI century documents. Several novel methodologies still under development and improvement were presented.

In what concerns the first two phases (page segmentation and separation of text from non-text) the morphological approach seems correct, presenting a high degree of generalisation as current test indicate.

Under the ultimate goal of word matching, strategies for lines and word segmentation were discussed and experimented. The obtained results can be considered very encouraging and they even will tend to improve, as further improvements are expected in the binarisation process.

The task of doing word recognition without the aid of the ‘semantic level’ is problematic. Humans and computers do not see symbols in the same way. Yet all the methods try to imitate more or less an imaginary process of how we recognise symbols. However, human brain performs a fusion of all the information captured by different senses. The same will have to be done here. Fuzzy partnership relations combined with neural networks decision rules, together with further developed

descriptors are expected to give improved results in the proposed task of word indexing in ancient books. This will be our goal as future work.

Acknowledgements

This research is being supported by the European Project LB5608/A – DEBORA (1999-2001), under the DGXIII-Telematics for Libraries Programme.

References

1. Agam G., Dinstein I., 1996, Adaptive Directional Morphology with Application to Document Analysis, in *Maragos P., Schafer R.W., Butt M.A. (eds.), Mathematical Morphology and its Applications to Image and Signal Processing*, 401-xxx, Kluwer Academic Publishers, Boston.
2. Beucher S., 1996, Pré-traitement morphologique d'images de plis postaux, 4^{ème} *Colloque National Sur L'écrit Et Le Document-Cned'96*, Nantes.
3. Bhat D., 1998, An Evolutionary Measure for Image Matching, in *ICPR'98 - Proc. 14th Int. Conf. On Pattern Recognition*, vol. I, 850-852, Brisbane, Australia.
4. Cinque L., Lombardi, L., Manzini G., 1998, A multiresolution approach to page segmentation, *Pattern Recognition Letters*, 19, pp 217-2225.
5. Cumplido M., Montolio P., Gasull A., 1996, Morphological Preprocessing and Binarization for OCR Systems, in *Maragos P., Schafer R.W., Butt M.A. (eds.), Mathematical Morphology and its Applications to Image and Signal Processing*, 393-400, Kluwer Academic Publishers, Boston.
6. Guillevic D., Suen C.Y., 1997, HMM Word Recognition Engine, in *ICDAR'97 - Proc. 4th Int. Conf. on Document Analysis and Recognition*, vol. 2, 544-547, Ulm, Germany
7. He S., Abe N., 1996, A Clustering-Based Approach to the Separation of Text Strings from Mixed Text/Graphics Documents, *Proceedings of ICPR '96*, Vienna.
8. Jain A.K., Yu B., Document Representation and its application to page decomposition, *IEEE Pattern Analysis and Machine Intelligence*, 20(3), pp 294-308, March 1998
9. Marcolino A., Ramos V., Ramalho M., Caldas Pinto J., 2000, Line and Word Matching in Old Documents, submitted to *SIARP'2000 - V Ibero-American Symposium on Pattern Recognition*, Lisboa.
10. Mengucci M., Granado I., Muge F., Caldas Pinto J.R., 2000, A Methodology Based on Mathematical Morphology for the Extraction of Text and Figures from Ancient Books, *RecPad 2000*, pp 471-476 Porto, 11-12 May 2000, Portugal.
11. Parodi P., Piccioli G., 1996, An Efficient Pre-Processing of Mixed-Content Document Images for OCR Systems, *Proceedings of ICPR '96*, Vienna.
12. Ramos V., 2000, An Evolutionary Measure for Image Matching – Extensions to Binary Image Matching, Internal Technical Report, CVRM/IST, Lisboa.
13. Serra J., 1982, *Image Analysis and Mathematical Morphology*, Academic Press, London.
14. Soille P., 1999, *Morphological Image Analysis*, Springer; Berlin.
15. Spitz A., 1999, Shape-based word Recognition, *International Journal on Document Analysis and Recognition*, vol 1, no. 4, 178-190.
16. Srihari, et al, *Document Image Understanding*, <http://www.cedar.buffalo.edu/Publications/TechReps/Survey/>, CEDAR-TR-92-1, 1992.
17. Tang Y.Y., Lee S.W., Suen C.Y., 1996, Automatic Document Processing: A survey; *Pattern Recognition*, 29(12), 1931-1952.

Content Based Indexing and Retrieval in a Digital Library of Arabic Scripts and Calligraphy

Suliman Al-Hawamdeh¹ and Gul N. Khan²

¹ School of Computer Engineering
Nanyang Technological University
Singapore 639798
Tel: (65) 790-5065, Fax: (65) 7926559
assuliman@ntu.edu.sg

² Department of Electrical Engineering
University of Saskatchewan
57 Campus Drive, Saskatoon, Saskatchewan, S7N 5A9
Canada

Abstract. Due the cursive nature of the Arabic scripts automatic recognition of keywords using computers is very difficult. Content based indexing using textual, graphical and visual information combined provides a more realistic and practical approach to the problem of indexing large collection of calligraphic material. Starting with low level pattern recognition and feature extraction techniques, graphical representations of the calligraphic material can be captured to form the low level indexing parameters. These parameters are then enhanced using textual and visual information provided by the users. Through visual feedback and visual interaction, recognized textual information can be used to enhance the indexing parameter and in return improve the retrieval of the calligraphic material. In this paper, we report an implementation of the system and show how visual feedback and visual interaction helps to improve the indexing parameters created using the low-level image feature extraction technologies.

1 Introduction

The flexibility and the artistic nature of the Arabic scripts have made it ideal for Arabic and Islamic Calligraphy. Calligraphy emerged as the most important medium in the Arab and Islamic world as Islam discouraged sculptural or figurative form of arts, which had its associations with idolatry. Today, there is a large collection of Arabic and Islamic Calligraphy available in digital format. A project was started two years ago to look into the technical issues

Name	Isolated	end	middle	start
Ba	ب	با	با	با
Ha	ح	حا	حا	حا
Seen	س	سا	سا	سا
A'in	ع	عا	عا	عا
Kaf	ك	كا	كا	كا
Dal	د	دا		
Ra	ر	را		

Fig. 1. Sample of Arabic Characters

related to the indexing and retrieval of large collection of Arabic and Islamic Calligraphy. As calligraphy is an artistic form of writing, words extracted using optical character recognition or pattern matching can be used as indexing parameters. Using low level feature extraction, it is possible to produce a set of features that can be used to form similar clusters for indexing. These clusters can then be enhanced using visual interaction and visual feedback. Unlike English and other Latin languages, Arabic is always written or printed cursively from right to left. Words consist of one or more connected portions and every letter has more than one shape depending on its position within the word. The portions forming the word can consist of one letter or multiple letters. The table in Figure 1 shows the Arabic alphabets and the different shapes depending on their position within the word. The discontinuities between different portions of the same words are due to some characters that cannot be connected to the letters, which come after them. These characters are shown in Figure 2. The flexibility and the artistic nature of the Arabic scripts have made it difficult for computer processing. There is a limited work and little progress in computer recognition of non-printed Arabic scripts and calligraphy [2,3]. Al-Muallim and Yamaguchi [2] discussed the problems associated with the handwritten scripts and showed examples of the cursive words and the complexity involved in separating the characters. Calligraphy on the other hand can take any shape and it is a more difficult and more complicated form of writing (Figure 3). Calligraphy was developed rapidly after the rise of Islam and used to beautify the words of God. It has been emerged as the most important medium in the Arabic and Islamic art.

With the increasing number of non-Arab Muslims and the increasing use of Arabic scripts in other languages, there was a need to adapt and reform the Arabic scripts. In Farsi or Persian for example four new characters are added to represent the phonetics that did not exist in Arabic. Other languages such as Turkish, Urdu, etc. have also added extra letters or dots to the basic Arabic scripts. In fact original Arabic script were limited and confusing since they share the same shape. As the need for reforming the letters arises, a new system of Nuqat (dots) and Tashkeel pointing was developed. This simplified the writing and reading of Arabic Scripts, but did not make it easier for computer processing.



Fig. 2. Letters that cannot be connected

In this paper, we report on the implementation of a digital library that contains a large collection of Arabic scripts and calligraphy. The initial indexing of Arabic scripts and calligraphy is carried out using feature extraction techniques. The indexing clusters generated by the low level feature extraction are then improved using visual feedback. While low-level feature extraction is carried out off-line, visual feedback is obtained online and in an interactive manner.

2 Image Features Extraction

Several image feature detection techniques have been investigated to extract features useful for indexing of Arabic calligraphic image. The basic and low-level features

that are applicable to Arabic calligraphic images are vertices, corners or high curvature points, short line and curved segments, etc. (Figure 4). The high-level image features used for indexing are the contours that can be easily formed by grouping low level features including edge points, short line segments, corners, curves, circles etc. An algorithm has been devised for identifying contour shapes based on the construction of a piece wise linear approximation of contours by straight and curved line segments and other feature points. The method involves the "perceptual" grouping of edge points after applying Hough transform for straight and curved line segment detection [9]. These line segments are linked into contours and during the linking process, straight line and curved segments and other feature points due to noise are filtered out. The filters used at this stage are also based on the perceptual criteria, and proved highly successful in extracting weak but perceptually significant contour. Generally, edge point detection is assumed to be a local and parallel process while the grouping has been considered as a global and sequential process. We have found that early and intermediate level grouping techniques, based on the proximity and similarity in orientation and brightness are useful for the extraction of contours from calligraphic images. It suggests that similarity in orientation; brightness and color are purely local relationships that can be used to devise grouping functions.



Fig. 3. Sample of Calligraphic material

Image contours provide plenty of clues about the calligraphic shapes. A number of constraints have been put forward for the interpretation of image contours. Some of these constraints determine the distance of a contour from other contours and they are used to determine similar calligraphic shapes.

2.1 Low Level Feature Detection

Main low-level features used for indexing calligraphic images are edge points in the form of vertices, corners and high curvature points, and line-segments both straight and curved ones. Edge points can be easily detected using well known edge detectors including the edge detector from Sobel, Laplacian, Prewitt and Canny. Corners and vertices are considered as second order derivatives and curvature extreme points. Corners can be detected either by enhancing the 2nd order gray level variation or by following the object boundary points and seek for local maximum curvature [11, 15]. Multi-resolution Fourier transform has also been used for detecting corner [6]. For straight line and curved segment detection Hough transform (HT) has been widely used. It was originally introduced by Hough [8], and later it was developed to group points into simple boundary features like straight lines and circular curves. It has also been generalized to extract complex shape features [4]. Hough techniques are robust in the presence of noise and they are not affected by missing information [14]. We

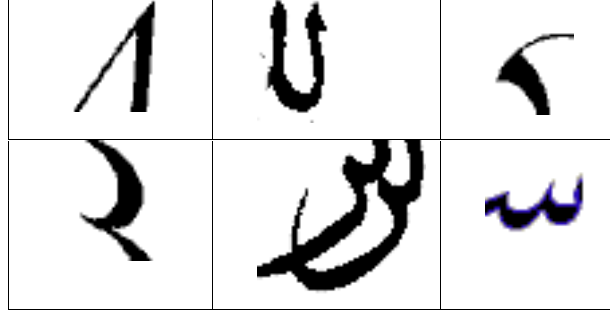


Fig. 4. Sample of the common features found in the calligraphic material

consider, binary calligraphic images consisting of edge points, which have been extracted in the initial stage of image data processing. For straight line extraction, HT approach transforms an edge point (x, y) to (ρ, θ) Hough space using the straight line equation

$$\rho = x \cos\theta + y \sin\theta$$

where ρ is the normal distance of line from the origin and θ denotes the angular position of normal as shown in Figure 5. By restricting θ in the interval $[-\pi/2, \pi/2]$, equation (1) transforms edge points in image space to sinusoidal curves in Hough space. The sinusoidal curves, which correspond to edge points on a particular line, have a common intersection point in the Hough space. An accumulator array, referred as Hough array is normally used to represent the discretized version of Hough space. Generally, the HT method has two computational phases: in the first phase, for each edge point (x_i, y_i) , ρ is calculated for all values of θ and the corresponding Hough array cell is incremented; in the second phase, Hough array cells are examined to find high count cells. These significant sets of collinear points belong to straight lines in the image.

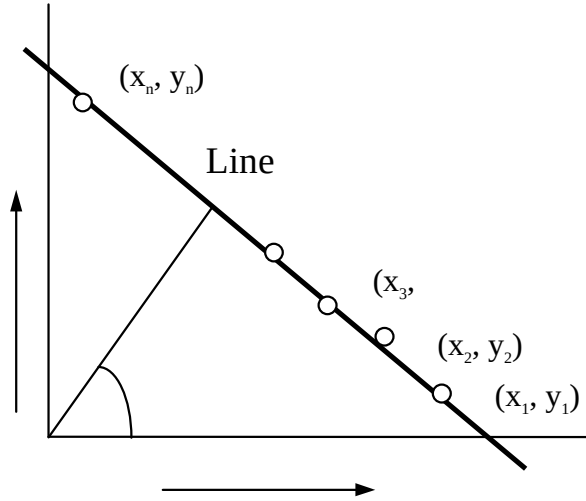


Fig. 5. Straight Line Representation in Hough space

To extract higher order boundary or shape features, a large dimensional Hough space is required. The circles and circular curved segments are detected by employing a 3D Hough space (x_c, y_c, r) based on the equation of a circle given below.

$$(x - x_c)^2 + (y - y_c)^2 = r^2$$

where (x_c, y_c) is the center and r is the radius of the circle or circular curve.

An improved algorithm based on the Dynamic Generalized Hough Transform, DGHT is employed to extract circular curves [16]. The computational complexity of HT grows exponentially with the higher order Hough space.

2.2 Contours Extraction by Grouping Line or Curved Segments

Most of the contour extraction techniques link edge points into contours directly without an intermediate representation. For extracting Arabic calligraphic image indexing features, We have adopted the strategy of forming a low-level feature representation first before grouping these features into contours. This has two advantages: firstly, it limits the use of domain specific knowledge during low-level feature extraction and at the early stages of visual data processing, and secondly it reduces the amount of data to be processed at higher levels. The lack of a simple signal to symbol paradigm for calligraphic images further support to the idea that a low-level feature representation should be formed as a first step in finding the higher-level contour representation. Many of the global curve detection techniques are based on graph search methods where the edges are viewed as the nodes of a graph, and a cost factor is associated with each link between nodes. The minimum cost paths in the graph are taken to correspond to the desired boundaries. Heuristics have also been applied to the search. Although this is basically a global method, an intermediate organization based on the streaks is used.

The edge pyramids have been used to extract boundaries of objects. A pyramid is constructed by reducing the resolution of the image at successive levels and then edge detectors are applied at each level. The edges between adjacent levels and at the same level are linked using proximity and orientation [7]. The extraction of smooth curves is carried out using an overlapped pyramid structure. In this method the curves are fed into the appropriate levels of the pyramid. The contours are approximated by line segments and at each level segments from the level below are combined using local position, curvature and direction.

Normally, in contour extraction techniques the edge point detection is assumed to be a local process and edge linking is carried out as a global and sequential process. The contour extraction method that can be used for Arabic calligraphic-images is based on an overlapped pyramid structure [10]. The boundary contours are extracted in a single pass, starting from the bottom level of the pyramids. Our approach utilizes a multi-resolution representation where short lines and curve segments, and other low-level features including corners or high curvature points and isolated clusters of edge points are fed into the pyramid from its leaves. The contours are constructed incrementally as the feature data travel upward from leaves to the root of the pyramid.

3 Free Text Indexing

Visual interaction and visual feedback make use of the captions and textual description assigned to images and multimedia information such as video and sound clips. Indexing of textual information is carried out using free text indexing techniques. Free text indexing used here employs a weighting function that has been shown to give better results than the other weighting techniques studied by Salton and Buckley [13]. The weighting function incorporates term frequency and length normalization. Selecting an image query that might consist of one or more images, the indexing terms correspond to these query images are extracted and used in formulating the textual query. When the search is carried out, similarity matching of retrieved images is calculated using the extracted terms from the query image and the index list in the database. The retrieved sets of images are then ranked in descending order according to their similarity with the image query.

Visual feedback refers to the subsequent searches in which a set of relevant images are identified by the user and used to expand the query. Al-Hawamdeh [1] showed that visual feedback increases the communication bandwidth between the user and the system so that the user needs can be more precisely specified, and improves the likelihood that the retrieved images will be relevant.

Using positive and negative feedback, query terms can be adjusted depending on their discriminant value and the degree to which they characterise relevant and no-relevant images. Price (1992) used document space modification techniques to improve the index generated by the short description of images in a picture archival system[12]. Quantitative tests of the modification algorithm were performed and the results showed that this represent a promising approach.



Fig. 6. User Interface with Arabic Support

The use of visual feedback and document space modification techniques provides a practical approach to the indexing problem. Rather than relying on users to describe visual and audio material or assign manually created headings, visual interaction and visual feedback provide a semi-automatic approach. Indexing terms and parameters can be extracted from the database and assigned to the new images based on their similarity to those exist in the database.

The system used in this study is designed to store, retrieve, view and manipulate wide range of image formats. It provides facilities for scanning, display, image

enhancements and manipulation utilities. Images can be stored in different formats including Tagged Image Formats (TIFF), Joint Photographic Group Formats (JPEG) and Graphic Image Format (GIF). Figure 5 shows the user interface of the system, which has a full Arabic support. The system allows the user to describe an image and use the image description to perform free text indexing.

4 Visual Feedback

Calligraphy is a type of art that can be converted into text for indexing and retrieval purposes. However, this conversion is not possible using currently available optical character recognition technology. Until today, Optical character recognition techniques for Arabic language are limited to printed letters with accuracy that does not exceed eighty percent in most cases. Recognizing the limitation of OCR, we have developed a new indexing technique based on low level feature extraction and visual feedback.

The feature extraction process is carried out off-line to determine the common features in the image and then determine the clusters in which the image can be classified under. The words in the calligraphic material are sometimes complex and it is not easy to recognize them automatically. Users familiar with calligraphic material might be able to understand the calligraphy and help interpreting the its content. The visual feedback process enables users to suggest a new clusters if they feel that the



Fig. 7. Shows the visual interaction and the evaluation method

calligraphic material was not properly indexed or it belongs to more than one cluster. Using relevance feedback and similarity matching, a new set of indexing parameters, whether keywords or extracted common features from images can be derived and used to enhance the indexing process.

To facilitate interaction with the users, a Web interface that allows users to enter a textual or graphical query was developed. Graphical queries are carried out using query by example in which one image is selected and used to find similar calligraphic images. The retrieved images are then ranked according to their similarity with the query. Figure 7 shows sample of the user interface in which multiple images are displayed on the screen along with the judgment criteria. Each image displayed on the screen has three database fields associated with it. If the user agrees with the results, then it means that the image retrieved is relevant and probably ranked satisfactory. If the user does not agree with the result, then the image might not be relevant or not ranked correctly. In this case the user can suggest alternative keywords. These keywords can be derived from the calligraphic material. When the user disagrees with the ranking, the system captures this feedback and depending on the user suggestions, it tries to adjust the indexing parameters.

5 Evaluation

To evaluate the effectiveness of visual feedback and visual interaction as a tool in improving the low level indexing using pattern-matching techniques, ten different queries were selected. In selecting the queries, we minimized similarities between queries to avoid repetition of suggested terms by the users. The five users participated

Queries	Initial Search	Second refinement	Third Refinement
1	30%	70%	80%
2	10%	50%	80%
3	30%	80%	100%
4	20%	80%	80%
5	50%	90%	100%
6	30%	70%	90%
7	10%	70%	80%
8	60%	80%	100%
9	30%	60%	70%
10	40%	80%	80%
Average	31%	73%	86%

Table 1. Shows the precision for the initial search as well as second and third refinement using visual feedback

in the experiment are selected on the basis of their familiarity with the Arabic calligraphy and their ability to judge the degree of relevance of the retrieved images to the initial query. The database used contains about 6000 calligraphic and manuscript images. The initial searches are refined using visual feedback and visual

interaction techniques by either including or excluding the image from the retrieved set. If the user disagree with the results and he did not offer any suggestion for indexing, then images are removed from the list and indexing parameters are modified accordingly. The test involved running ten initial searches and thirty subsequent refined searches. Results of initial searches based on feature extraction alone showed poor results. Users on average agreed with only three images out of the top 10 images displayed. A cut off point of 10 is used in which the top 10 most similar images to the query are displayed. The average precision for the initial searches across the ten queries used was 31%. After the second refinements, the average precision was improved to 73%. This happened after excluding the non-relevant images from the hit list and also after modifying the indexing parameters according to the alternative indexing terms given by the user. The average precision for the third refinement, which includes changes from initial and second refinement, was increased to 86%. Again this improvement is mainly due to excluding the non-relevant images and improving the rank for those images in which the users provided alternative terms. One interesting observation is that the results in the second refinement was improved for some of the queries in which indexing parameters were modified by other users. As non-relevant images were excluded from one of the query hits, they appeared as relevant on the other query hit lists. This is due to the alternative indexing terms provided by the users. The results listed in table 1 showed that visual feedback and visual interaction could significantly improve the indexing of calligraphic material. While automatic pattern recognition can be used to create a low level clusters of calligraphic material, visual feedback and users interaction can be used improve the indexing and retrieval process.

6 Conclusion

Content based indexing of calligraphic material using combined graphical, textual, and visual information provides a practical approach to the indexing of large number of calligraphic material. Using low level pattern recognition and feature extraction techniques, it is possible to create some basic clusters of similar images based on similar features found in the calligraphic images. The classification of these images might not be accurate at this stage but by using visual feedback and visual interaction we can improve the indexing process by tapping users skills and knowledge about the calligraphic material. Obvious concern when dealing with users over the Internet is accuracy of information provided. Despite this concern, results showed that visual feedback and visual interaction could provide practical and cheaper solution to the problem of indexing and retrieval of a large collection of calligraphic images.

References

1. A. Al-Hawamdeh et. al. Nearest Neighbour searching in a picture archival system. Proceeding of the ACM International Conference on Multimedia and Information Systems. Singapore, Jan. 16-19, (1991) 17-33.

2. H. Al-Muallim, S. Yamaguchi A Method of recognition of Arabic cursive handwriting. IEEE Transaction on Pattern Analysis and Machine Intelligence. 9 (1987) 715-722.
3. A. Amin, H. Al-Sadoun and S. Fischer. (1996) Hand-printed Arabic character recognition system using an artificial network. Pattern Recognition. 29 (1996) 663-675.
4. D. H. Ballard Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition*, 13 (2) (1981) 111-122.
5. W. B. Croft, D. J. Harper. (1979) Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*. 35 (1979) 285-295.
6. A. R. Davies, and R. Wilson (1992) Curve and Corner Extraction using the Multiresolution Fourier Transform, International Conference on Image Processing and its Applications, (2992) pp. 282-285.
7. T. Hong, M. Shneier, and A. Rosenfeld (1982) Border Extraction using Linked Edge Pyramids. IEEE Transaction on System Management and Cybernetics, 12(5) (1982) 631-635.
8. P.V.C. Hough *Method and means for recognizing complex patterns*, U.S. Patent (1962) No. 3069654.
9. G. N. Khan, D. F. Gillies (1992) Extracting contours by perceptual grouping, *Image and Vision Computing*, 10 (2) (1992) 77-88.
10. G. N. Khan, D. F. Gillies A parallel-hierarchical method for grouping line segments into contours', SPIE's Proceedings of the 33rd International Symposium, Application of Digital Image Processing XII, San Diego, California (1989) pp. 237-246.
11. J. Alison Nobel (1988) Finding corners. *Image and Vision Computing*. 6, (1988) 121-128.
12. R. Price, T. S. Chua, S. Al-Hawamdeh. (1992) Applying relevance feedback to a photo archival system. *Journal of Information Science*, 8,(1992) 203-215.
13. G. Salton, C. Buckley Term weighting approaches in automatic text retrieval. *Information Processing and Management*. 24 (1989) 513-523.
14. K. Varsha, S. Ganesan (1998) A Robust Hough Transform technique for Description of Multiple Line Segments in an Image", (4-7 October) International Conference on Image Processing, 1 (1998) 216 - 220 (ISBN: 0-8186-8821-1)
15. W. Wehnu, A. V. Jose Segmentation of Planar Curves into Straight-line Segments and Elliptical Arcs. *Graphical Models and Image Processing*, 59(6) (1997) 484-494.
16. Y. Wei, Circle Detection using Improved Generalized Hough Transform (IDHT). (6-10 July) IEEE International Conference on Geoscience and Remote Sensing Symposium IGARSS Vol.2 (1998) 1190-1192, (ISBN: 0-7803-4403-0)

Ancient Music Recovery for Digital Libraries

J. Caldas Pinto *, P. Vieira *, M. Ramalho*, M. Mengucci **, P. Pina **, and F. Muge**

* IDMEC/IST - Technical University of Lisbon - Instituto Superior Técnico,
Av. Rovisco Pais, 1049-001 Lisboa, PORTUGAL

** CVRM/Centro de Geo-Sistemas - Instituto Superior Técnico
Av. Rovisco Pais, 1049-001 Lisboa, PORTUGAL

Abstract. The purpose of this paper is to present a description and current state of the “ROMA” (*Reconhecimento Óptico de Música Antiga* or Ancient Music Optical Recognition) Project that consists on building an application, for the recognition and restoration specialised in ancient music manuscripts (from XVI to XVIII century). This project, beyond the inventory of the *Biblioteca Geral da Universidade de Coimbra* musical funds aims to develop algorithms for scores restoration and musical symbols recognition in order to allow a suitable representation and restoration on digital format. Both objectives have an intrinsic research nature one in the area of musicology and other in digital libraries.

Introduction

The generality of National and Foreign libraries suffer from the problem of having to keep and administrate rich collection of musical manuscripts mainly from centuries XVI to XIX. Much of that music was composed by important composers. Nowadays we report a growing interest in those manuscripts, either by musicologists or Ancient Music performers both nationals and foreigners. However, libraries have two main problems to tackle: documents preservation and edition. Nowadays libraries are impotent against its collections’ degradation. The reading of the scores is becoming increasingly impossible, due to continuous process of degradation of the physical documents and internal (paper, ink, etc.) and external (light radiation, heat, pollution) factors. On the other hand it is frequently necessary to photocopy those manuscripts being that also a harmful operation. This project has a twofold objective: inventorying the *Biblioteca Geral da Universidade de Coimbra* musical funds and developing algorithms for scores restoration and musical symbols recognition, naturally for a suitable and significant class of scores. Both objectives have an intrinsic research nature one in the area of musicology and other in digital libraries.

Statement of the Problem

ROMA (Ancient Music Optical Recognition) is a project intended to recover ancient music scores manuscripts to obtain a digital, easy to manage and easy to conserve,

and, last but not the least, easy to handle music heritage. Optical Music Recognition is the process of identifying music from an image of a music score. The music scores under consideration are, most of the times, on paper. Identifying music means constructing a model of the music printed, in some format that enables the score to be re-printed on paper or even played by a computer. This formats capture the semantics of music (notes pitches and times) instead of the image of a music score, bringing, among others, the following advantages: (i) It occupies considerably less space; (ii) It can be printed over and over again, without loss of quality; (iii) It can be easily edited (with a proper editor). These advantages will bring self-correction capabilities to the system under development.

OMR has some similarities with OCR (Optical Character Recognition). In OMR, instead of discovering what character is in the image, the aim is to discover what musical symbol is in the image (including notes, rests, clefs, accidents, etc.). However it can not be supported by a dictionary although some grammatical rules can aid as support for misunderstood signs.

Ancient music recognition raises additional difficulties as: (i) notation varies from school to school and even from composer (or copyist) to composer (or copyist) (ii) simple (but important) and even sometimes large changes in notation occur in the same score; (iii) staff lines are mostly not the same height, and not always are straight; (iv) symbols were written with different sizes, shapes and intensities; (v) the relative size between different components of a musical symbol can vary; As some documents were hand written additionally (i) more symbols are superimposed in hand-written music than in printed music; (ii) different symbols can appear connected to each other, and the same musical symbol can appear in separated components; (iii) paper degradation requires specialised image cleaning algorithms. Maybe because of these difficulties, attempts to tackle this problem are sparse in literature ([1], [5]).

Basic Process of OMR

The most common and simple approach to OMR, found in most literature, is through a set of stages composed in a pipelined architecture: the system works from stage to stage, from the first to the last, each stage producing results that are the inputs of the next stage. The most common stages are:

- **Pre-processing of the image:** corrections at the level of image processing to simplify the recognition. Includes image rotation, to straighten staff lines, binarization (transforming a coloured image to a black and white one), image cleaning (in ROMA, this is a key issue: degradation of the paper requires great investment in image cleaning).
- **Removal of non-musical symbols.** This stage consists of removing symbols that are not relevant to the music.
- **Identification (and removal) of staff lines:** localisation of the lines in the image and deletion, to obtain a new image having only musical objects ([1], [2] and [5]).
- **Object segmentation.** This is the process of identification (recognition) of simple objects like blobs or lines that make part of a musical symbol, isolating them and constructing an internal model of the objects ([6]).

- **Object reconstruction.** The stage on which the isolated simple objects are assembled in all sorts of musical symbols (notes, rests, etc.). For this process it is usual to use DCG's (Defined Clause Grammars) that describe each musical symbol through its components [9].
- **Constructing the final representation.** The final stage performs a transformation of the identified musical objects in a musical description format, such as NIFF or MIDI ([10]).

The work that has already been done in this project is related to the stages of identification and removal of staff lines and object segmentation. We present the developments next.

Image Pre-processing

A pre-processing approach mainly based on mathematical morphology operators [10][11] was developed.

Input: The input is constituted by true colour images of approximately 2100x1500 pixels. They are constituted by lyrics and music scores printed in black, over a light yellow background, decorated with blue, red and gold signs like “illuminated letters” (figure 1a) or by particular notations within the music scores (figure 1b). In addition, the background of the images reveals the printed signs of the verso (other side of the pages) with lighter intensity, apart from other normal dirt resulting from natural causes or human handling.

Output: The segmentation/classification of the several different components of the initial coloured images will produce binary images.

Steps: The developed algorithm is constituted by 2 steps described in the following: (1) segmentation of coloured signs, (2) segmentation of music scores.

1. Segmentation of Coloured Signs

The colour images are converted from Red-Green-Blue (RGB) to Hue-Intensity-Saturation (HIS) colour spaces, better permitting to classify the coloured signs and to extract separately important marks, which are simply black signs (notes and lyrics, staff lines).

Once the coloured signs present strong colours, its segmentation is quite simple and mainly based on the combination of simple thresholdings on the Hue and Saturation channels. The sets corresponding to the previously identified main colours (red, gold-yellow and blue) are separated into three different binary images. The application of these masks to the images of figure 1, is shown in figure 2.



Fig. 1. Coloured signs in musical scores

This information is very useful in the next step when it will be necessary to make a distinction of the dark signs to segment from some the coloured surplus.

2. Segmentation of Music Scores

Due to the colourless aspect of the music scores, its segmentation is performed on the image of the channel Intensity.

After the application of a smoothing filter (median) to remove local noise, the morphological gradient is used as the basis to construct a mask by dilation, that covers all the significant structures, *i.e.*, the most contrasting structures in relation to the background. Within this binary mask the segmentation of the music scores is obtained through the reconstructed gradient approach developed to segment pages in books of the Renaissance [12].

Once the application of the reconstructed gradient approach is limited to the mask zone, most of noise far located from the music scores is suppressed.



Fig. 2. Segmentation of colored signs: (a) gold-yellow, and (b) red

The application of the developed methodology to a full page coloured page (figure 3a) is presented in figures 3b and 3c, respectively for the music scores and for the coloured signs.

Identification and Removal of Staff Lines

Indeed a method for identifying staff lines is already implemented in a musical manuscript from the XVI century with promising results. The algorithm is as follows:

Input: a binary image having an unknown number of staff lines.

Output: a coloured image, with the staff lines marked with a different colour. (Not only the centre of the line is marked, but also all the black pixels identified as belonging to staff lines are coloured. The pixels belonging to other objects that superimpose with the staff lines are not coloured.

Steps: The general algorithm has 2 steps: (1) Finding the staff lines, and (2) Marking the staff lines. These are described next.

Note: For this objective (detecting the staff lines), a simple Hough transform technique could have been used. However, we find our method more adaptable and extensible in regard to future work, in staff images where lines are not completely straight (having slight curvatures) and have different heights across the image (see [1]).

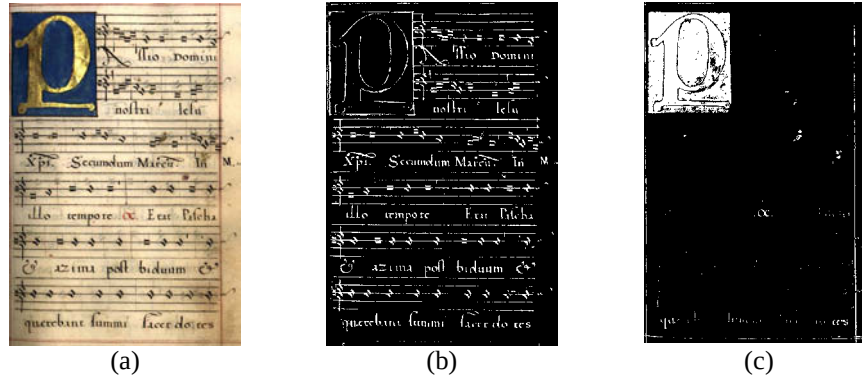


Fig. 3. (a) Full coloured page of a musical score; (b) Segmented music scores after subtraction of coloured signs; (c) Segmented coloured signs with vivid colour (blue, red, yellow)

1. Finding the Staff Lines

The first step consists of finding the centre of each line belonging to a staff line. This is done by analysing the horizontal projection of the black pixels. The horizontal projection is the counting of the number of black pixels, for each line of the image. A perfect line of a staff should correspond to a line of the image that has only black pixels. We consider that the lines that have the biggest count of black pixels (the greater projection) belong to staff lines. We define a threshold, which we call **projection threshold**, for which the projections, that are greater than that threshold, are considered staff lines. How do we find this threshold? We explain it later. It matters now to see that, around the staff lines, there will be areas of lines whose projection exceeds the projection threshold. We mark the centre of these areas as the centre of the lines. (Figure 4)

The decision process is as follows:

For all l_1, l_2 (where l_1 and l_2 are lines of the image, and $l_2 > l_1$), If for all l , such as $l_1 < l < l_2$, $\text{projection}(l) > \text{projection\_threshold}$, Then $c = l_1 + ((l_2 - l_1) / 2)$ is the centre position of a line.

Rotating the Image. To find the angle for which the staff lines in the image are horizontal, the algorithm calculates the horizontal projection for the image, rotated between two pre-defined angles, α_1 and α_2 , with a pre-defined interval. The best angle, α_0 , will be the one in which the centre of the horizontal lines has the greater projection (number of pixels):

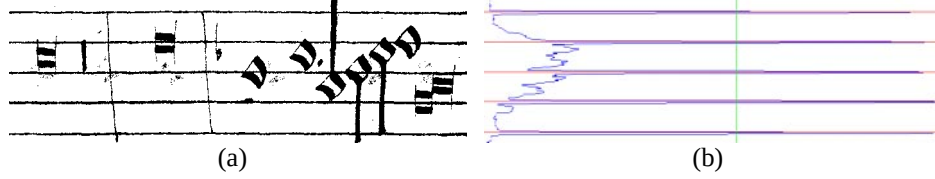


Fig. 4. A clip from a binary image of a hand-written music score. (b) - Horizontal projection at blue, best projection threshold at green and centre of the max areas at red (from a)

$\alpha_b = \arg \max_{\alpha} \{ \text{MaxSum}(\alpha, \text{Image}) \}$, where $\alpha_1 < \alpha < \alpha_2$, and $\text{MaxSum}(\alpha, \text{Image}) = \sum \text{projection}(l)$, where l is a line of *Image* where $\text{projection}(l)$ is a local max of the projection, and $\text{projection}(l) > \text{projection\_threshold}$.

(See figure 5, and compare it with figure 4: notice that the projection of figure 5 is smaller).

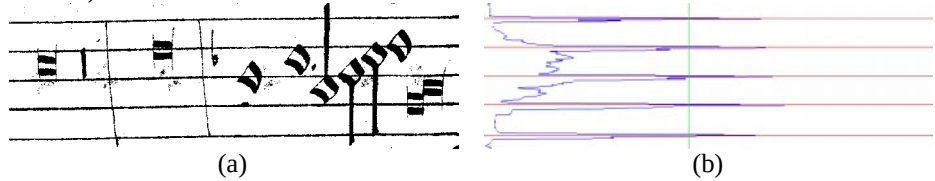


Fig. 5. A clip from a binary image of a hand-written music score, with a rotation of 0.5 degrees right. (b) 2b - horizontal projection at blue, best projection threshold at green and centre of the max areas at red (from figure 5a)

Finding the Projection Threshold. We use two different projection thresholds along the algorithm. For the issue of rotating the image, we use a pre-defined threshold. This value is fixed because it is not too relevant for finding the best angle. 40% is the best value for the images currently being handled. For the issue of finding the staff lines themselves, we execute an algorithm that finds the best projection threshold for the image. The algorithm is as follows:

Input: The horizontal projection of the image:

Output: An integer (corresponding to the best projection threshold)

Process: Starting from the pre-defined threshold, the algorithm incrementally tests several thresholds above and below the initial one, to find the best one. For each threshold, it counts the number of staff lines in the image (a staff line is a group of 5 lines with similar distance between them; note that there may be some lines in the image that do not belong to staff lines). The best threshold is the one for which the number of lines belonging to staff lines minus the number of all lines is maximum.

$\text{best\_projection\_threshold} = \arg \max_{\text{threshold}} \{ \# \text{StaffLines}(L) - \#L \}$, where L is the set of lines found in the horizontal projection, using a certain *threshold* (see the decision process above), and $\text{StaffLines}(L)$ is a subset of L , corresponding to the lines that belong to staff lines.

2. Marking the Staff Lines

After we have found the centre of each line in a staff line, we will mark all the pixels that belong to the line. Note that we want to separate the pixels that belong to lines

from the pixels that belong to musical objects. A straightforward approach is followed (e.g. [6]). It works as follows:

Inputs: The image, and the positions of the centre of the lines belonging to staff lines.

Outputs: The image, with the lines marked.

Steps: For each line,

1. Estimate the line width.
2. For all columns of the image,
 - 2.1. Retrieve the black stripe of that column (set of black pixels that are vertically connected to the pixel of the centre of line).
 - 2.2. Mark or not the black stripe, according to following decision process:
 If $\text{length}(\text{black\_stripe}) < \text{stripe\_threshold}$, MARK, Else, DON'T MARK, where, $\text{stripe\_threshold} = \text{stripe\_threshold\_factor} * \text{estimated\_line\_width}$

(The stripe threshold is proportional to the expected width of the lines of the staff lines, since we want to remove the lines but not the objects that superimpose with them. The stripe threshold factor was calculated empirically and has the value of 1.6. However, an estimate or learning process should be used in the future for application to other pieces of music. More sophisticated decision processes could have been used (for instance, a template matching process with the objects that superimpose with the lines could identify more clearly the pixels that belong to lines and the ones that belong to objects). However, our simple process has proven to work finely for the images currently being handled.)

The final result is shown on figure 6.

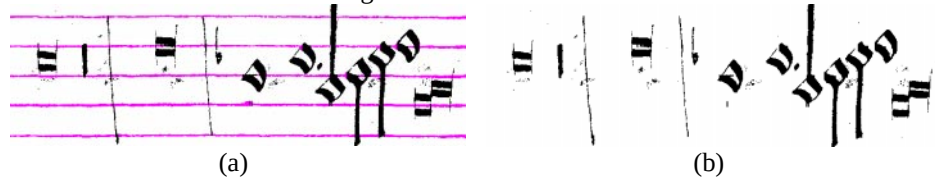


Fig. 6. Lines of the staff marked with pink (from the image on figure 6a). (b) - Lines of the staff deleted (from the image on figure 6a).

Object Segmentation

Object segmentation is the process of identifying and separating different components of the image that represent musical objects or parts of musical objects. In ROMA, a pre-segmentation process exists, to identify bar lines. In fact, this is the completed work regarding object segmentation, in this project.

Identifying Bar Lines

Described OMR systems (e.g., [5]), the bar lines are recognised after the segmentation process. In ROMA targeted images it is more efficient to know the bar line positions prior to this process, because (1) The bar lines are visibly different from the rest of the objects in the image, and (2) They divide the score in smaller regions,

producing block segmentation and the definition of regions of interest as a decomposition of the image [4].

The bar line identification process can be seen as a 2-phased classification algorithm. Here it is:

Input: The image, The line boundaries of the staff line

Output: Centre positions (columns of the image) of the bar lines

Steps:

1. Calculate the vertical projection of the staff line (see figure 7).

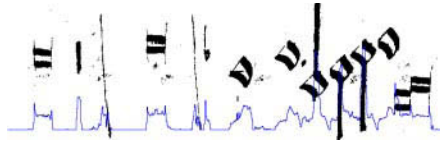


Fig. 7. Vertical projection at blue (from the image on figure 6b).

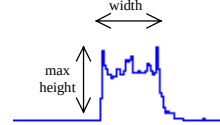


Fig. 8. Features of a local maxima window: max height and width.

2. Pick local maxima regions (windows) where all the vertical projection in the window is above a given threshold (**window height threshold**)

$$W = \{ (c1, c2) : \forall c, c1 \leq c \leq c2, \text{vertical\_projection}(c) > \text{window\_height\_threshold} \}$$

3. First classification process: choose bar line candidates from the local maxima regions of W . The domain is W and the feature vector is $f^w = (\text{width}, \text{max\_height})$, where, for a pair $(c1, c2)$ belonging to W , $\text{width} = c2 - c1$, which is the width, in pixels, of the window above the window height threshold, and $\text{max\_height} = \max_c \{ \text{vertical\_projection}(c) \}$, for $c1 < c < c2$, which is the maximum value of the vertical projection in the window (see figure 8).

The classification process is as follows:

Let $w \in W$, and $f^w = (wd, hg)$ be the feature vector of w , w is a bar line candidate if: (1) $wd1 < wd < wd2$, and (2) $hg > hg1$, where $wd1$, $wd2$ and $hg1$ are pre-defined.

(The values of $wd1$, $wd2$ and $hg1$ were set empirically, and should be estimated or automatically learned in the future, for application to other pieces of music. This process seems to be enough, as the following rules apply: (1) The area, in pixels, of the bar lines are mostly identical, (2) Bar lines are always almost vertical, and (3) Bar lines are isolated from other objects, in the image.)

Some other musical objects can be included as bar lines candidates, when we use the above features (see figure 9a). To differentiate them, we use an additional feature in a second phase classification, which is the *standard deviation* of the horizontal projection of the window (see figure 9b).

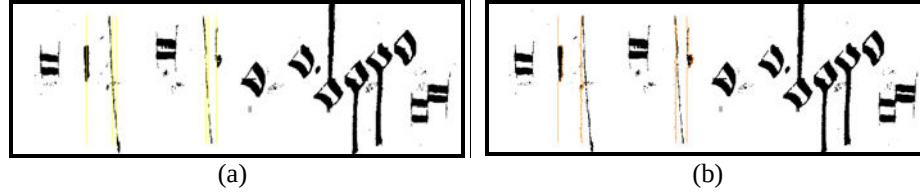


Fig. 9. (a) - Several objects, besides bar lines, are classified as bar line candidates. (b) - Horizontal projection of each window classified as a bar line candidate.

4. Second classification process: choose the bar lines, from the bar line candidates.

Let $w \in \text{bar\_line\_candidates}(W)$, and $f^w = (sd)$ be the feature vector of w , where $sd = \text{standard\_deviation}(\text{horizontal\_projection}(w))$, w is a bar line window if: $sd > sd1$, where $sd1$ is pre-defined.

(The value of $sd1$ was set empirically, but should be estimated or automatically learned in the future, for application to other pieces of music. This process seems to be enough, since the bar lines are the only objects that occupy the whole height of the staff line and, at the same time, pass through the prior classification process.)

Experimental Results

We present next the final result for a full page of the musical piece from the XVI century that is being used for experiments (Figure 10).

Conclusions and Future Work

The completed work in our project is a system that can mark (or remove) the staff lines from an image of a music score (possibly slightly rotated), and locate the bar lines in the score. Currently, the system has been tested for a set of images from a XVI century musical piece. Effectively there is a set of parameters that were set empirically, and for their generalisation and calculus a more extensive analysis should be performed.

Future work includes not only the automatic calculation or estimation of the referred parameters, but also the continuing process of music recognition. These would require the completion of the object segmentation, recognition and assembling procedures, and the translation of the results to a musical representation language, or the printing of the recognised music.

Acknowledgements

This project is supported by the Portuguese National Project PRAXIS/C/EEI/12122/1998.

et: & filius hominis tra-

detur ut crucifigatur. & Tunc

congregati sunt principes sa-

cerdotu, & seniores po-

puli, in atrium principis

sacerdotu, qui dicebatur

Fig. 10. Full page of a musical piece from the XVI century, with the staff lines removed and the bar lines located.

References

1. F. Pépin, R. Randriamahefa, C. Fluhr, S. Philipp and J. P. Cocquerez, "Printed Music Recognition", IEEE - 2nd ICDAR, 1993
2. I. Leplumey, J. Camillerapp and G. Lorette, "A Robust Detector for Music Staves", IEEE - 2nd ICDAR, 1993
3. K. C. Ng and R. D. Boyle, "Segmentation of Music Primitives", British Machine Vision Conference, 1992
4. S. N. Srilhari, S. W. Lam, V. Govindaraju, R. K. Srilhari and J. J. Hull, "Document Image Understanding", Center of Excellence for Document Analysis and Recognition
5. Kia C. Ng and Roger D. Boyle, "Recognition and reconstruction of primitives in music scores", Image and Vision computing, 1994
6. David Bainbridge and Tim Bell, "An Extensible Optical Music Recognition system", Ninetenth Australian Computer Science Conference, 1996
7. "Musical notation codes: SMDL, NIFF, DARMS, GUIDO, abc, MusiXML", http://www.s-line.de/homepages/gerd_castan/compmus/notationformats_e.html
8. Christian Pellegrini, Mélanie Hilario and Marc Vuilleumier Stückelberg, "An Architecture for Musical Score Recognition using High-Level Domain Knowledge", IEEE - 4^o ICDAR, 1997
9. Jean-Pierre Armand, "Musical Score Recognition: A Hierarchical and Recursive Approach", IEEE - 2^o ICDAR, 1993
10. Jean Serra, "Image Analysis and Mathematical Morphology", Academic Press, London, 1982.
11. Pierre Soille, "Morphological Image Operators, Springer", Berlin, 1999.
12. Michele Mengucci and Isabel Granado, "Morphological Segmentation of Text and Figures in Renaissance Books (XVI Century)", to present at ISMM'2000 – 5th International Symposium on Mathematical Morphology and its Applications to Image and Signal Processing, Palo Alto, USA, June 2000.

Probabilistic Automaton Model for Fuzzy English-Text Retrieval

Manabu Ohta, Atsuhiro Takasu, and Jun Adachi

National Institute of Informatics (NII),
Hitotsubashi 2-1-2, Chiyoda-ku, Tokyo 101-8430, Japan,
{ohta, takasu, adachi}@nii.ac.jp

Abstract. Optical character reader (OCR) misrecognition is a serious problem when searching against OCR-scanned documents in databases such as digital libraries. This paper proposes fuzzy retrieval methods for English text that contains errors in the recognized text without correcting the errors manually. Costs are thereby reduced. The proposed methods generate multiple search terms for each input query term based on probabilistic automata reflecting both error-occurrence probabilities and character-connection probabilities. Experimental results of test-set retrieval indicate that one of the proposed methods improves the recall rate from 95.56% to 97.88% at the cost of a decrease in precision rate from 100.00% to 95.52% with 20 expanded search terms.

1 Introduction

Automatic document image processing systems [10], which convert printed documents into their digital forms such as Standard Generalized Markup Language (SGML) documents, have been extensively researched as a result of improvements in both the optical character reader (OCR) and document image analysis. On the other hand, recognition errors that the OCR process inevitably produces must be dealt with for a search of the OCR output text [2]. Correcting errors manually is a conventional approach, however, fuzzy retrieval methods for noisy text have recently been proposed because of the high cost of manual post-editing. In other words, OCR-scanned raw text without error corrections or with only an automatic, not manual, correction by a Post-Processing System (PPS)[6] if any, has recently been stored in digital libraries mainly accumulating document images.

For example, the text retrieval group at the Information Science Research Institute (ISRI) at the University of Nevada, Las Vegas, developed a document processing system that facilitates the task of constructing the functional electronic forms of document collections[11]. This system has several remarkable features to automate the data construction process, one of which is an automatic document markup system[10] to automatically mark words, sentences, paragraphs, and sections in OCR-scanned documents and generate Standard Generalized Markup Language (SGML) documents. They also extensively researched the effects of OCR errors on text retrieval[7,9,2,8] and concluded that

for a simple boolean retrieval system, the problems caused by OCR errors can be overcome by exploiting the natural redundancy in the document text. However, they also indicated that the influence of OCR errors was not negligible for retrieval of shorter, less redundant documents or for more sophisticated and complicated retrieval systems such as ranking retrieved documents or adopting relevance feedback.

This paper, therefore, proposes fuzzy retrieval methods[3,4,5] specifically designed to handle such OCR-scanned text and shows that good retrieval effectiveness is achieved in the English-text retrieval of words¹ in spite of existence of errors. The proposed methods do not correct any errors included in the OCR-scanned text but make allowance for them in the retrieval process. These methods are based on probabilistic automata in which the states are assigned one or two original characters and the output symbols are assigned one or two non-original characters. In addition, possible recognition results of the original characters assigned to a state are given as an output symbol at the transition to the state[1]. The proposed methods generate multiple search terms by chaining all the output symbols when the automata accept a query term, which makes up for retrieval omissions caused by OCR errors. The proposed methods, of course, can be applied to post-processed text, because PPS can be considered as part of the OCR function.

2 Proposed Methods

Fuzzy retrieval methods proposed in this paper expand a query term α into plural search terms $\{\beta_1, \beta_2, \dots, \beta_n\}$ by referring to error-occurrence and character-connection probabilities, which are integrated in the probabilistic automata constructed beforehand by using a training set. All the generated search terms are ranked according to the calculated probability $P(\beta_i|\alpha)$, which represents whether β_i is appropriate as an expanded term for the original query term α . During the retrieval process, the top n expanded terms, β_i , are used instead of α to compensate for retrieval omissions caused by OCR errors.

In this paper, OCR errors are classified into five categories: i) substitution, ii) deletion, iii) insertion, iv) combination, and v) decomposition errors because these categories covered over 97% of English OCR errors in our preliminary experiments[5]. Therefore, an OCR is assumed to recognize a character correctly unless it produces one of these five errors. A substitution error is defined to have $(1, 1)$ correspondence, a deletion error has $(1, 0)$, an insertion error has $(0, 1)$, a combination error has $(2, 1)$, and a decomposition error has $(1, 2)$, where l and m are the length of an original and OCR output character string, respectively, and (l, m) represents the correspondence of both strings. The other types of errors are assumed to be composed of the defined five types of errors, e.g. (l, l) errors composed of l substitution errors, $(l, 0)$ errors composed of l deletion errors, and so on.

¹ The proposed methods search for a word (term), not for a document.

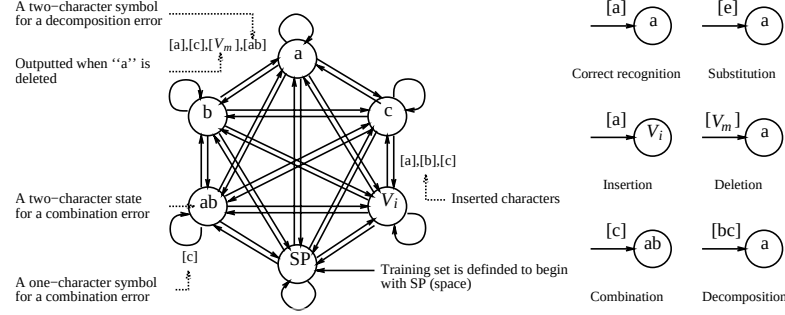


Fig. 1. A simple example of probabilistic automata and OCR's behavior on them

2.1 Probabilistic Automata Construction

The probabilistic automata for expanding a query term, shown in Fig. 1, are constructed beforehand. The states of these automata correspond to original characters, and the output symbols correspond to non-original (OCR generated) characters. These states consist of one character, two contiguous characters, and V_i , which is a virtual character that corresponds to an inserted character. In addition, the output symbols consist of one character, two contiguous characters, and V_m , which is a virtual character that corresponds to a deleted character. The state transition probabilities represent character-connection probabilities based on character bigram statistics. The symbol output probabilities represent the recognition probabilities that depend on the OCR employed.

The behavior of the OCR on the probabilistic automata is explained as follows. The recognition of x as y by the OCR invokes a transition to the state x and at the same time outputs the symbol y . Therefore, the transition to a one-character state is invoked when an original character is recognized correctly, substituted, deleted, or decomposed. In addition, the transition to a two-character state is invoked when two contiguous original characters are combined and outputted as one non-original character. The transition to the state V_i is invoked when a non-original character is inserted during the OCR process. At the state transition, one-character, two-character, or a V_m symbol is also outputted, depending on the kind of misrecognition by the OCR. In particular, substitution, insertion and combination errors, in addition to correct recognition, invoke an output of a one-character symbol, deletion errors invoke that of the symbol V_m because OCR does not output a character for a deletion error, and decomposition errors invoke that of a two-character symbol into which an original character is decomposed.

2.2 State Transition and Symbol Output Probabilities Estimation

An original text of sufficient length and its recognition result are necessary as a training set for constructing probabilistic automata. A comparison of the non-

[Original character|Output character]
[t][t][h][h][i][i][s][s][|][i][i][s][s][|][a][o][|][s][s][a][e][m][m][p][p][l][l][e][e][.][.]

Fig. 2. Tagged text for training probabilistic automata

original text with its original text and some heuristics for extracting and categorizing recognition errors produce the text shown in Fig. 2 where correspondence between original and non-original (OCR generated) characters have been determined explicitly. The text shown in Fig. 2, makes it possible to calculate state transition and symbol output probabilities by counting the frequency of state transitions and symbol outputs.

Let $C(s_i \rightarrow s_j)$ be the number of times the transition from states s_i to s_j is used in a training set. The transition probability is defined as follows.

$$p_{s_i s_j} = \frac{C(s_i \rightarrow s_j)}{\sum_j C(s_i \rightarrow s_j)}. \quad (1)$$

On the other hand, the symbol output probability is defined in two different ways. Let $C(s_i \xrightarrow{sym_k} s_j)$ be the number of times the transition from states s_i to s_j and the output of the symbol sym_k are used in a training set. In addition, let the symbol output probability on the probabilistic automaton of type A (PA of type A) be defined as follows.

$$q_{s_i s_j}(sym_k) = \frac{C(s_i \xrightarrow{sym_k} s_j)}{\sum_l C(s_i \xrightarrow{sym_l} s_j)}. \quad (2)$$

Moreover, let the symbol output probability on the probabilistic automaton of type B (PA of type B) be defined as follows.

$$q_{s_i s_j}(sym_k) = q_{s_j}(sym_k) = \frac{\sum_i C(s_i \xrightarrow{sym_k} s_j)}{\sum_i \sum_l C(s_i \xrightarrow{sym_l} s_j)}. \quad (3)$$

Please note that the symbol output probabilities on the PA of type A are bound to the state transition, whereas those of type B are bound to the state.

2.3 Generation of State Sequences

All possible state transition sequences for an input query term α are generated according to the following rules.

- The initial state is defined as any delimiting character such as a space and period.
- The second state is either a state of the first character of a query term or a two-character state consisting of its first and second characters when such a two-character state exists.

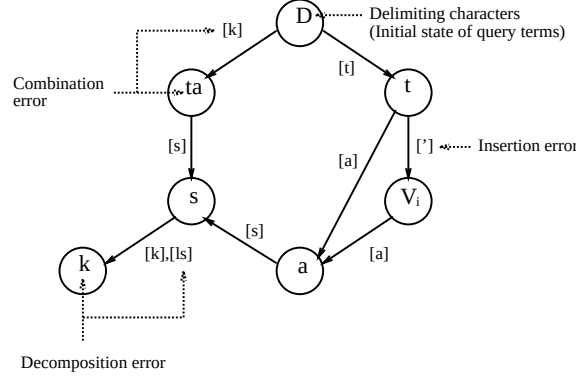


Fig. 3. Part of probabilistic automata extracted for a query term “task”

- The third or later state is the state of the next character of the previous state character, a two-character state (if a combination error occurs), or a state V_i (if an insertion error occurs).
- The possibility of the occurrence of more than two contiguous insertion errors and any insertion errors before and after a query term is ignored.

For example, Fig. 3 shows part of constructed probabilistic automata relevant to a query term “task”. In Fig. 3, arrows are drawn at the only state transitions whose probabilities are relatively large and that compose “task”. Therefore, the following three state sequences for the query term “task” are acquired according to the above rules.

1. State sequence 1: “D” → “t” → “a” → “s” → “k”
2. State sequence 2: “D” → “ta” → “s” → “k”
3. State sequence 3: “D” → “t” → “ V_i ” → “a” → “s” → “k”

Let $S_x^\alpha = s^1 s^2 \dots s^n$ be one of the state transition sequences acquired by the above rules. $P(S_x^\alpha)$ can be calculated as follows

$$P(S_x^\alpha) = \pi_{s^1} \prod_{i=1}^{n-1} p_{s^i s^{i+1}} = \prod_{i=1}^{n-1} p_{s^i s^{i+1}}. \quad (4)$$

where the initial state probability, π_{s^1} , is 1 because the initial state is defined to be the state of a set of delimiters. When all the state sequences and the probabilities of their occurrence have been acquired in this way, the probability $P(S_x^\alpha | \alpha)$, based on the probabilistic automata, is calculated as follows.

$$P(S_x^\alpha | \alpha) = \frac{P(S_x^\alpha)}{P(\alpha)} = \frac{P(S_x^\alpha)}{\sum_x P(S_x^\alpha)}. \quad (5)$$

The sum of the probabilities of the finite number of state sequences, S_x^α , approximates $P(\alpha)$ in (5) due to the restriction of not allowing contiguous insertion errors.

2.4 Output Symbol Sequences

For a given state sequence, S^α , the symbol sequences SYM^β that depend on it are outputted with the probability $P(SYM^\beta|S^\alpha)$, which is calculated using (6).

$$P(SYM^\beta|S^\alpha) = \prod_{i=1}^{n-1} q_{s^i s^{i+1}}(sym^{i+1}). \quad (6)$$

To take an example of “task” shown in Fig. 3, the following six symbol sequences are outputted depending on the three state sequences enumerated in Sect. 2.3.

1. State sequence 1: [t][a][s][k], [t][a][s][ls]
2. State sequence 2: [k][s][k], [k][s][ls]
3. State sequence 3: [t]['] [a][s][k], [t]['] [a][s][ls]

2.5 Score of the Expanded Terms

When a symbol sequence is given, the probability $P(SYM^\beta, S_x^\alpha|\alpha)$ that a query term α chooses a state sequence S_x^α and outputs a symbol sequence SYM^β is estimated to be the product of $P(S_x^\alpha|\alpha)$ (cf. (5)) and $P(SYM^\beta|S_x^\alpha)$ (cf. (6)). The probability $P(SYM^\beta|\alpha)$ that a query term α outputs a symbol sequence SYM^β irrespective of its state sequence is estimated to be the sum of $P(SYM^\beta, S_x^\alpha|\alpha)$.

$$P(SYM^\beta|\alpha) = \sum_x P(SYM^\beta|S_x^\alpha)P(S_x^\alpha|\alpha). \quad (7)$$

Each output symbol sequence SYM_y^β is then converted into its corresponding character string β_z . Please note that $P(\beta_z|SYM_y^\beta) = 1$ holds true while $P(SYM_y^\beta|\beta_z) = 1$ generally does not hold. Therefore, $P(\beta_z|\alpha)$, on which the ranking of expanded terms is based, is calculated as follows.

$$P(\beta_z|\alpha) = \sum_y P(\beta_z|SYM_y^\beta)P(SYM_y^\beta|\alpha). \quad (8)$$

3 Experiments

3.1 Outline

The method performance was experimentally evaluated in terms of both retrieval effectiveness and the number of expanded search terms. The retrieval effectiveness refers here to the recall (the ratio of appropriate results retrieved by the proposed method to actual character strings satisfying the input query) and precision (the ratio of appropriate results retrieved by the proposed method to all results retrieved by the proposed method) rates in terms of a full-text search for strings. The experiments were carried out on text data of about 650

kilobytes collected from *Artificial Intelligence and Information Processing and Management*, 1995 to 1998², obtained from Elsevier Electronic Subscriptions. In the experiments, five-fold cross-validation was adopted, that is, the text data were divided into five sets, one of which was a test set that had nothing to do with making probabilistic automata whereas the others composed a training set that was used to make them. The experiments were done five times varying the combination of training and test sets because there were five different combinations of them. We randomly selected 50 nouns appearing in all the five test sets at least five times as query terms³. Plurals were distinguished from their singulars. The retrieval effectiveness values and the number of expanded search terms are average values of the five experiments as well as of the retrieval of 50 terms. The other experimental conditions are as follows.

1. All uppercase characters were transformed into the corresponding lowercase characters beforehand (normalization). Words were extracted based on 19 types of delimiters such as a space and period.
2. The OCR system employed was found to have recognition accuracy of 99.01% after the above normalization.
3. All the transition probabilities whose values were estimated at 0 by training were assigned 0.00001 (flooring).
4. Both state transition and symbol output probabilities concerning the transitions from all the two-character states were smoothed by using those from one-character states which were identical to the second character of the two-character states. Such smoothing by using one-character states was introduced because combination errors producing two-character states did not occur frequently and, as a consequence, these probabilities were hard to be estimated.

3.2 Results

The resultant retrieval effectiveness and the number of expanded search terms are summarized in Tables 1 and 2. The recall (R) and Precision (P) rates were $R = 95.70\%$ and $P = 100.00\%$ in the training-set retrieval, and $R = 95.56\%$ and $P = 100.00\%$ in the test-set retrieval when noisy text was searched by an original query term without any query term expansions, that is, exact matching, which is considered to be a bottom line. On the other hand, Figs. 4 and 5 show the retrieval effectiveness for searching the training and test sets as a function of the number of expanded terms (the expanded terms with the highest scores were used).

Figure 4 for training-set retrieval indicates that the proposed method based on the probabilistic automaton of type A (PA A method) achieved $R = 98.73\%$ and $P = 94.23\%$ and the method based on the probabilistic automaton of type

² Strictly speaking, the text data collected from *Information Processing and Management*, Vol.35, No.1, 1999 was also included.

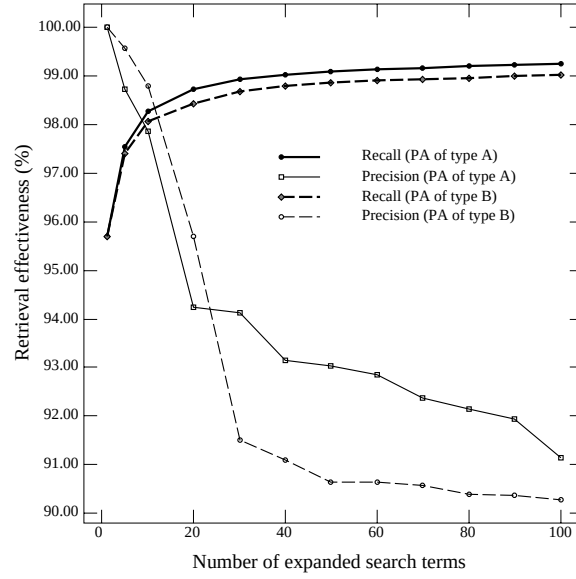
³ Some examples of the query terms are “algorithm”, “information”, and “system”.

Table 1. Retrieval effectiveness when searching the training set

Retrieval condition	Recall(%)	Precision(%)	Number of terms
Exact matching	95.70	100.00	1
PA A method	98.73	94.23	20
PA B method	98.42	95.70	20

Table 2. Retrieval effectiveness when searching the test set

Retrieval condition	Recall(%)	Precision(%)	Number of terms
Exact matching	95.56	100.00	1
PA A method	97.86	94.17	20
PA B method	97.88	95.52	20

**Fig. 4.** Retrieval effectiveness vs. the number of expanded terms in the training-set retrieval

B (PA B method) achieved $R = 98.42\%$ and $P = 95.70\%$ with 20 expanded terms. Both methods gained recall rates of about 3% at a cost of about 6% (PA A method) or about 4% (PA B method) decrease in precision rates. On the whole, the PA A method outperformed the PA B method in recall rates. In precision rates, however, the PA B method outperformed the PA A method with

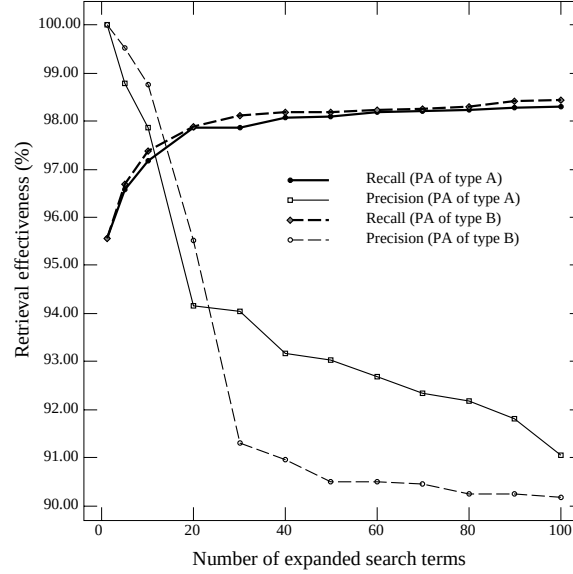


Fig. 5. Retrieval effectiveness vs. the number of expanded terms in the test-set retrieval

smaller number of expanded terms (up to about 20 terms) while the situation was reversed with larger number of expanded terms.

Figure 5 for test-set retrieval indicates, with 20 expanded terms, the PA A method achieved $R = 97.86\%$ and $P = 94.17\%$ while the PA B method achieved $R = 97.88\%$ and $P = 95.52\%$. Both methods gained recall rates of about 2.3% at a cost of about 6% (PA A method) or about 4.5% (PA B method) decrease in precision rates. On the whole, the PA B method outperformed the PA A method in recall rates, which is in sharp contrast to the results shown in Fig. 4. In precision rates, the test-set retrieval showed a tendency similar to the training-set retrieval depending on the number of expanded terms.

4 Conclusions

This paper proposed fuzzy retrieval methods for noisy English text. These methods expand a query term into multiple search terms based on the probabilistic automata. Their significant feature is that the expansion process naturally reflects both recognition characteristics of the employed OCR (symbol output probabilities) and language characteristics on character connections (state transition probabilities).

The experimental results indicated that with 20 expanded terms, one of the proposed methods gained about 2.3% of recall rates at a cost of about 4.5% decrease in precision rates in the test-set retrieval. The experimental results also indicated that binding symbol output probabilities to state transitions (PA of

type A) was preferable to binding them to the states (PA of type B) in the training-set retrieval. The latter, however, outperformed the former in recall rates for the test-set retrieval because the size of the training set used was insufficient for parameters of the PA of type A to be estimated accurately enough. Consequently, the PA B method performed better due to the smaller number of parameters. Therefore, the trade-off between the size of a training set and the retrieval effectiveness needs to be examined. Optimal selection of the probabilistic automata according to the size of a training set will also be considered in our future work.

References

- [1] Eugene Charniak. *Statistical Language Learning*. The MIT Press, 1993.
- [2] W. B. Croft, S. M. Harding, K. Taghva, and J. Borsack. An evaluation of information retrieval accuracy with simulated OCR output. In *Proc. of SDAIR'94 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 115–126, Las Vegas, NV, April 1994.
- [3] Daniel Lopresti and Jiangying Zhou. Retrieval strategies for noisy text. In *Proc. of SDAIR'96 5th Annual Symposium on Document Analysis and Information Retrieval*, pages 255–269, Las Vegas, NV, April 1996.
- [4] Daniel P. Lopresti. Robust retrieval of noisy text. In *Proc. of ADL'96 Forum on Research and Technology Advances in Digital Libraries*, pages 76–85, Library of Congress, Washington, D. C., May 1996. URL <http://dlt.gsfc.nasa.gov/adl96/>.
- [5] Manabu Ohta, Atsuhiko Takasu, and Jun Adachi. Reduction of expanded search terms for fuzzy English-text retrieval. In *Proc. of ECDL'98, LNCS 1513*, pages 619–633, Crete, Greece, September 1998. Springer.
- [6] Kazem Taghva, Julie Borsack, and Allen Condit. An expert system for automatically correcting OCR output. In *Proc. of the IS&T/SPIE 1994 International Symposium on Electronic Imaging Science and Technology*, pages 270–278, San Jose, CA, February 1994.
- [7] Kazem Taghva, Julie Borsack, and Allen Condit. Results of applying probabilistic IR to OCR text. In *Proc. of the 17th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pages 202–211, Dublin, Ireland, July 1994.
- [8] Kazem Taghva, Julie Borsack, and Allen Condit. Effects of OCR errors on ranking and feedback using the vector space model. *Information Processing & Management*, 32(3):317–327, 1996.
- [9] Kazem Taghva, Julie Borsack, and Allen Condit. Evaluation of model-based retrieval effectiveness with OCR text. *ACM Trans. on Information Systems*, 14(1):64–93, January 1996.
- [10] Kazem Taghva, Allen Condit, and Julie Borsack. An evaluation of an automatic markup system. In *Proc. of the IS&T/SPIE 1995 International Symposium on Electronic Imaging Science and Technology*, pages 317–327, San Jose, CA, February 1995.
- [11] Kazem Taghva, Allen Condit, Julie Borsack, John Kilburg, Changshi Wu, and Jeff Gilbreth. The MANICURE document processing system. Technical Report 95-02, Information Science Research Institute, University of Nevada, Las Vegas, NV, March 1995.

Associative and Spatial Relationships in Thesaurus-Based Retrieval

Harith Alani¹, Christopher Jones², and Douglas Tudhope¹

¹School of Computing, University of Glamorgan, Pontypridd, CF37 1DL, UK
{halani, dstudhope}@glam.ac.uk

²Department of Computer Science, Cardiff University, Cardiff, CF24 3XF, UK
C.B.Jones@cs.cf.ac.uk

Abstract. The OASIS (Ontologically Augmented Spatial Information System) project explores terminology systems for thematic and spatial access in digital library applications. A prototype implementation uses data from the Royal Commission on the Ancient and Historical Monuments of Scotland, together with the Getty AAT and TGN thesauri. This paper describes its integrated spatial and thematic schema and discusses novel approaches to the application of thesauri in spatial and thematic semantic distance measures. Semantic distance measures can underpin interactive and automatic query expansion techniques by ranking lists of candidate terms. We first illustrate how hierarchical spatial relationships can be used to provide more flexible retrieval for queries incorporating place names in applications employing online gazetteers and geographical thesauri. We then employ a set of experimental scenarios to investigate key issues affecting use of the associative (RT) thesaurus relationships in semantic distance measures. Previous work has noted the potential of RTs in thesaurus search aids but the problem of increased noise in result sets has been emphasised. Specialising RTs allows the possibility of dynamically linking RT type to query context. Results presented in this paper demonstrate the potential for filtering on the context of the RT link and on subtypes of RT relationships.

1 Introduction

Recent years have seen convergence of work in digital libraries, museums and archives with a view to resource discovery and widening access to digital collections. Various projects are following standards-based approaches building upon terminology and knowledge organisation systems. Concurrently, within the web community, there has been growing interest in vocabulary-based techniques, with the realisation of the challenges posed by web searching and retrieval applications. This has manifested itself in metadata initiatives, such as Dublin Core and the proposed W3C Resource Description Framework. In order to support retrieval, provision is made in such metadata element sets for thematic keywords from vocabulary tools such as thesauri (ISO 2788, ISO 5964). Metadata schema (ontologies) incorporating thesauri or related semantic models underpin diverse ongoing projects in remote access, quality-based services, cross domain searching, semantic interoperability, building RDF models and digital libraries generally ([5], [10], [15], [29]).

Thesauri define semantic relationships between index terms [3]. The three main relationships are Equivalence (equivalent terms), Hierarchical (broader/narrower terms: BT/NTs), Associative (Related Terms: RTs) and their specialisations. A large number of thesauri exist, covering a variety of subject domains, for example MEDical Subject Headings and the Art and Architecture Thesaurus [2]. Various studies have supported the use of thesauri in online retrieval and potential for combining free text and controlled vocabulary approaches [16]. However there are various research challenges before fully utilising thesaurus structure in retrieval. In particular, the ‘vocabulary problem’ [10], differences in choice of index term at different times by indexers and searchers, poses problems for work in cross domain searching and retrieval generally. For example, indexer and searcher may be operating at different levels of specificity, and at different times an indexer(s) may make different choices from a set of possible term options. While conventional narrower term expansion may help in some situations, a more systematic approach to thesaurus term expansion has the potential to improve recall in such situations. In this project, we have employed the Getty AAT [38] and TGN (Thesaurus of Geographic Names) [20] vocabularies. Harpring [21] gives an overview of the Getty’s vocabularies with examples of their use in web retrieval interfaces and collection management systems. It is suggested that the AAT’s RT relationships may be helpful to a user exploring topics around an information need and the issue of how to perform query expansion without generating too large a result set is also raised.

The work described here is part of a larger project, OASIS (Ontologically Augmented Spatial Information System), exploring terminology systems for thematic and spatial access in digital library applications. One of our aims concerns the retrieval potential of geographical metadata schema, consisting of rich place name data but with locational data limited to a parsimonious approximation of spatial extent, or footprint. Such geographical representations may be appropriate for online gazetteers, geographical thesauri or geographic name servers, where conventional GIS datasets are unavailable, unnecessary or pose undesirable bandwidth limitations [22]. Notable projects include the Alexandria Digital Library [17]. Another aim explores the potential of reasoning over semantic relationships to assist retrieval from terminology systems. Measures of semantic distance make possible imprecise matching between query and information item, or between two information items, rather than relying on an exact match of terms [42]. Previous work investigated hybrid query/navigation tools based on semantic closeness measures over the purely hierarchical Social History and Industrial Classification [14]. This paper describes an integrated spatial and thematic schema and discusses two novel approaches to the application of thesauri, from both spatial and thematic points of view.

In section 2 we discuss our schema, illustrating how the spatial relationships in the thesaurus can be used to provide more flexible retrieval for queries incorporating place names. The second topic (sections 3 and 4) concerns the use of associative thesaurus relationships in retrieval. Existing collection management systems include access to thesauri for cataloguing with fairly rudimentary use of thesauri in retrieval (mostly limited to interactive query expansion/refinement and Narrower Term expansion). In particular, there is scope for increased use of associative (RT) relationships in thesaurus-based retrieval tools. RTs are non-hierarchical and are sometimes seen as weaker relationships. There is a danger that incorporating RTs into retrieval tools with automatic query expansion may lead to excessive ‘noise’ being introduced into result sets. We discuss results from scenarios with semantic distance

measures in order to map key issues affecting use of RTs in retrieval. Conclusions are outlined in section 5.

2 OASIS Overview and Spatial Access Example

Thematic data was taken mainly from the Royal Commission on the Ancient and Historical Monuments of Scotland (RCAHMS) database, which contains information on Scottish archaeological sites and historical buildings [31] – see Figure 2 for an example. The OASIS ontology was linked to the AAT which provided thematic descriptors such as ‘town’, ‘arrow’, ‘bronze’, ‘axe’, ‘castle’, etc. The spatial data in the OASIS system includes information on hierarchical and adjacency relations between named places, in addition to place types, and (centroid) co-ordinates. This information was taken from the TGN, augmented with data derived from the Bartholomew’s [19] digital map data for Scotland.

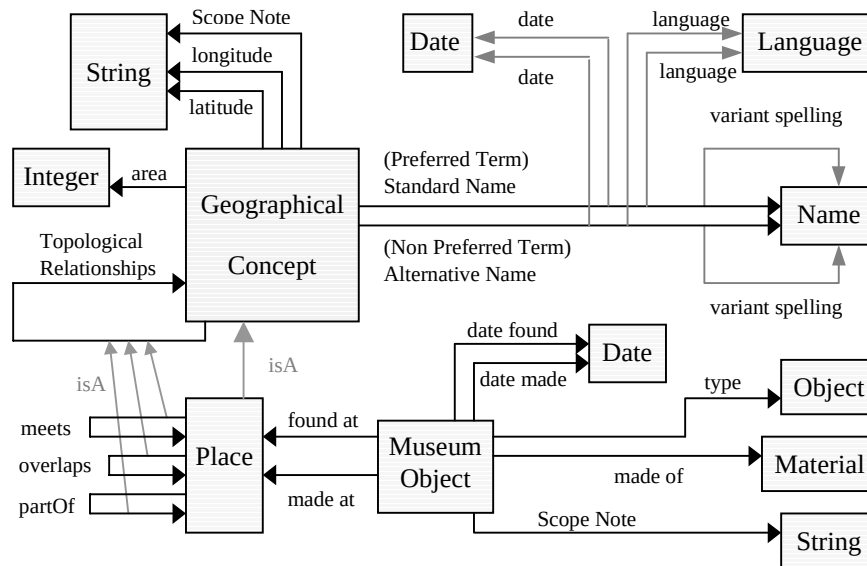


Fig. 1. The Classification schema of *Place* and *Museum Object* in the OASIS system.

The term ‘ontology’ has widely differing uses in different domains [18]. Our usage here follows [5] in viewing an ontology as a conceptualisation of a domain, in effect providing a connecting semantics between thesaurus hierarchies with specifications of roles for combining thesaurus elements. The OASIS schema (Figure 1) encompasses different versions of place names (e.g. current and historical names, different spellings), place types (e.g. Town, Building, River, Hill), latitude and longitude co-ordinates, and topological relationships (e.g. meets, part of). The schema is implemented using the object-oriented Semantic Index System (SIS [12]) also used to store the data, and which provided the AAT implementation. The SIS has a meta modelling capability and an application interface for querying the schema. Figure 1

shows the meta level classification of the classes *Place* and *Museum Object*. As we discuss later in relation with RTs, relationships can be instantiated or subclassed from other relationships. Thus, *meets*, *overlaps*, and *partOf* are subclasses of *Topological Relationships*. The relationships *Standard Name* and *Alternative Name* are instances of the relationships *Preferred Term* and *Non Preferred Term* respectively (shown in brackets). The *variant spelling* relationship links the place name (standard or alternative) to its spelling variations. *Place* inherits relationships, such as *longitude* and *latitude*, from its superclass, *Geographical Concept*. The information stored in the OASIS database can be accessed via a set of functions through which it is possible to find information related to a given place, or find objects at a place made of a certain material, etc. For example, to find all places within the City of Edinburgh, the system returns a set of all the places linked with a *partOf* relationship to the City of Edinburgh.

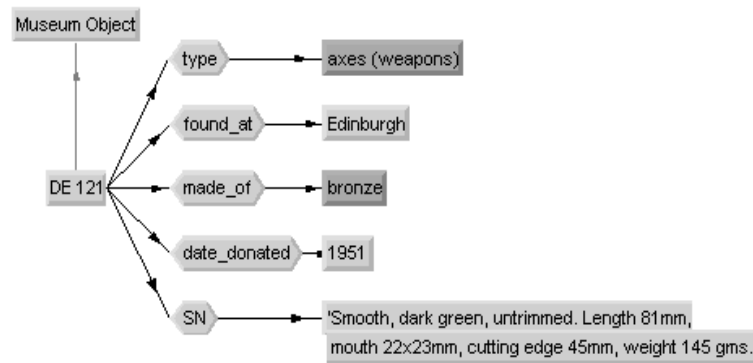


Fig. 2. Classification of the axe artefact NMRS Acc. No. DE 121.

Figure 2 shows the OASIS classification of an axe artefact from the RCAHMS dataset. OASIS implements a set of thematic and spatial measures that enables query expansion to find similar terms. Consider the query *Do you have any information on axes found in the vicinity of Edinburgh?*. An exact match to the query would only return axes indexed by the term *Edinburgh* (as in Figure 2). To search for axes found in the vicinity, spatial distance measures can expand the geographical term *Edinburgh* to spatially similar places, where axes have been found. Conventional GIS measures can be applied in situations where a full GIS polygon dataset is available. However, there are contexts where a GIS is either not appropriate (due to lack of co-ordinate data or bandwidth limitations) or where qualitative spatial relationships are important, eg remote access to online gazetteers and application contexts where administrative boundaries are important [22].

In our database, a query on axe finds would return several places, including *Carlops*, *Corstorphine*, *Harlow Muir*, *Hermiston*, *Leith*, *Penicuik*, *Tynehead*, *West Linton*. These places can be ranked by spatial similarity using the *Part-of* spatial containment relationship, which in OASIS is based on the spatial hierarchies in the TGN. Given the term *Edinburgh*, the OASIS spatial hierarchy distance measure ranks

Corstorphine, Leith, Tynehead equally and ahead of the other places listed, since (like *Edinburgh*) they are districts within the region *City of Edinburgh*. Similarly, since *Carlops* etc are places in Scotland, they would be returned ahead of any axe finds in England. In fact, the TGN provides centroid co-ordinate data for places/regions and our larger project explores the integration of different spatial distance measures and boundary approximation methods, based on geographical thesaurus relationships and limited locational footprint data [4].

3 Semantic Distance Measures

A thesaurus can be used as a search aid to a user constructing a query by providing a set of controlled terms that can be browsed via some form of hypertext representation (eg [7], [33]). This can assist the user to understand the context of a concept, how it is used in a particular thesaurus and provide feedback on number of postings for terms (or combinations of terms). The inclusion of semantic relationships in the index space, moreover, provides the opportunity for knowledge-based approaches where the system takes a more active role by reasoning over the relationships [42]. Candidate terms can be suggested for user consideration in refining a query and various forms of automatic query expansion are possible. For example, information items indexed by terms semantically close to query terms can be returned in a ranked result list. Imprecise matching between two media item is also possible in 'More like this item' options. The various Okapi experiments [6] investigated the extent to which thesauri should play an interactive or automatic role in query expansion.

The basis for such automatic term expansion is some kind of semantic distance measure. Semantic distance between two terms is often based on the minimum number of semantic relationships that must be traversed in order to connect the terms [34]. Each traversal has an associated cost factor. In poly-hierarchical systems, variations have been based on common or uncommon superclasses ([36], [39], [40]), or have employed spreading activation ([9], [11], [13], [32]). Rada et al [34] assigned an identical cost to each traversal, whereas other work has assigned different weights depending on the relationship involved ([28], [25], [27]). Sometimes depth within the hierarchical index space has been a factor, with distance between two connected terms considered greater towards the top of a hierarchy than towards the bottom, based on arguments concerning relative specificity, density or importance ([36], [39]). Other issues include similarity coefficients between sets of index terms ([37], [41]). Our focus in this paper is upon factors particularly relevant to the use of RTs in retrieval.

RTs represent a class of non-hierarchical relationships, which have been less clearly understood in thesaurus construction and applicability to retrieval than the hierarchical relationships. At one extreme, an RT is sometimes taken to represent nothing more than an extremely vague 'See-also' connection between two concepts. This can lead to an introduction of excessive noise in result sets when RT relationships are expanded. Rada et al [34] argue from plausible demonstration scenarios that semantic distance measures over RT relationships can be less reliable than over hierarchical relationships, unless the user's query can be closely linked to the RT relationship - a medical expert system example is given in [35]. As we discuss later, structured definitions of RTs (eg [3]) offer potential for systematic approaches

to their use. There is some evidence that RTs can be useful in retrieval situations. The basic assumption of a cognitive basis for a semantic distance effect over thesaurus terms has been investigated by Brooks [8], in a series of experiments exploring the relevance relationships between bibliographic records and topical subject descriptors. These studies employed the ERIC database and thesaurus and consisted of purely linear hierarchies, as opposed to tree hierarchical structures (as with the AAT) or indeed poly-hierarchies. However the results are suggestive of the existence of some semantic distance effect, with an inverse correlation between semantic distance and relevance assessment, dependant on position in the subject hierarchy, direction of term traversal and other factors. In particular, a definite effect was observed for RTs (typically less than for hierarchical traversal). An empirical study by Kristensen [26] compared single-step automatic query expansion of synonym, narrower-term, related term, combined union expansion and no expansion of thesaurus relationships. Thesaurus expansion was found to improve recall significantly at some (lesser) cost in precision. Taken separately, single step RT expansion results did not differ significantly from NT or synonym expansion (specific results showing a 12% increase in Recall over NTs, but with 2.8% decrease in Precision). In another empirical study by Jones [24], a log was kept of users' choices of relationships interactively expanded via thesaurus navigation while entering a query. In this study of users refining a query, a majority of terms retrieved from the thesaurus came from RTs (the then INSPEC thesaurus contained many more RTs than hierarchical relationships).

4 RT Scenarios and Discussion

This section maps key issues affecting use of RTs in term expansion. Results are given from a series of scenarios applying different versions of a semantic distance algorithm to terms in the AAT [2]. The distance measure employed a branch and bound algorithm, with weights for relationships given below and a depth factor which reduced costs according to hierarchical depth. It was implemented in C++ using the SIS function library to query the underlying schema given in Figure 1.

Our aim was to investigate different factors relevant to RT expansion, rather than relative weighting of relationships. In general the purpose of weighting relationships is to achieve a ranking in 'semantically close' terms to allow a user to either choose a candidate term to expand a query or to select an information item from a result set deriving from an automatic query expansion. When assigning weights to relationships it should be noted that there may be a dependency on type of application and particular thesaurus involved. The choice of threshold to truncate expansion is an associated factor, which may in practice be made contingent on some user indication of the degree of flexibility desired in results. The weights chosen for this experiment were selected to reflect some broad consensus of previous work. Commercial collection management or retrieval systems employing a thesaurus tend to be restricted to narrower term expansion (if any), thus favouring NTs. McMath et al [28] assigned costs of 10, 15 to NT and BT respectively. Chen and Dhar [9] employed weights of 9, 5, and 1 for NT, RT, and BT relationships respectively. Their weights were set according to the use frequency of relationships during empirical search experiments. Cohen and Kjeldsen's [11] spreading activation algorithm traversed NT before BT. The weights employed here (BT 3, NT 3, RT 4), taken together with a

depth factor inversely proportional to the hierarchical depth of the destination term, assign lowest costs to NTs and favour RTs over BTs at higher depths in the hierarchy (following an AAT editorial observation that RTs appear to work better at fairly broad levels). The threshold used to terminate expansion was 2.5.

We developed a series of experimental scenarios based around term generalisation involving RT traversal. Building on the example in Section 2, we focus on the AAT's *Objects Facet: Weapons & Ammunition* and *Tools & Equipment* hierarchies. The AAT, a large, evolving thesaurus widely used in the cultural heritage community, is organised into 7 facets, with 33 hierarchies as subdivisions, according to semantic role. The introductory scenario supposes a narrowly defined information need for items concerning axes used as weapons - mapping to AAT term *Axes (weapons)*. In the initial scenario, let us suppose expansion is restricted to NT relationships only. This yields: *tomahawks (weapons)*, *battle-axes*, *throwing axes*, and *franciscas*.

The second scenario supposes an information need for items more broadly associated with axes used as weapons. We first consider expansion over hierarchical relationships. Table 1 shows results from hierarchical (BT/NT) expansion only.

Table 1. BT/NT expansion only. Semantic distance shown for each term.

Term	Dist.	Term	Dist.	Term	Dist.
axes (weapons)	0	halberds	2.35	poniards	2.35
tomahawks	0.6	pollaxes	2.35	stiletos (daggers)	2.35
battle-axes	0.6	gisarmes	2.35	trench knives	2.35
edged weapons	1	bills (staff weapons)	2.35	arm daggers	2.35
throwing axes	1.1	corsescas	2.35	fighting bracelets	2.35
franciscas	1.53	glaives	2.35	finger hooks	2.35
staff weapons	1.75	integral bayonets	2.35	finger knives	2.35
sword sticks	1.75	knife bayonets	2.35	brass knuckles	2.35
harpoons	1.75	plug bayonets	2.35	switchblade knives	2.35
bayonets	1.75	socket bayonets	2.35	dirks	2.35
daggers (weapons)	1.75	sword bayonets	2.35	bolos (weapons)	2.35
fist weapons	1.75	left-hand daggers	2.35	bowie knives	2.35
knives (weapons)	1.75	cinquedeas	2.35	Landsknecht daggers	2.35
swords	1.75	ballock daggers	2.35	<swords by form>	2.35
partisans	2.35	baselards	2.35	<swords by function>	2.35
spears (weapons)	2.35	eared daggers	2.35	weapons	2.5
leading staffs	2.35				

Table 2 shows the effect of introducing RT expansion¹. Note that staff weapons related to axes are brought now closer (*halberds*, *pollaxes*, *gisarmes*) and new terms, (such as *axes (tools)*, *chip axes*, *ceremonial axes*) are introduced. The latter set of terms could be relevant to broader information needs or to situations when a thesaurus entry term was mismatched (in this case the information need might relate more to tool use). In some situations however, the RTs could be seen as noise.

¹ When term expansion is extended to RTs in a distance measure including a depth factor, it becomes important to base RT depth on the starting (not destination) term. Otherwise, two terms one link away could appear at different distances if they came from different hierarchical levels and this distortion is propagated to subsequent BT/NT expansions.

Table 2. RT expansion included.

Term	Dist.	Term	Dist.	Term	Dist.
axes (weapons)	0	adze-hatchets	1.9	sword bayonets	2.35
tomahawks (weapons)	0.6	hewing hatchets	1.9	left-hand daggers	2.35
battle-axes	0.6	lathing hatchets	1.9	cinquedeas	2.35
edged weapons	1	shingling hatchets	1.9	ballock daggers	2.35
axes (tools)	1	<cutting tools>	2	baselards	2.35
halberds	1	fascies	2	eared daggers	2.35
pollaxes	1	Pulaskis	2	(Landsknecht	
gisarmes	1	(<ceremonial		daggers)	2.35
ceremonial axes	1	weapons>)	2	poniards	2.35
throwing axes	1.1	(<wood-cutting		stiletos (daggers)	2.35
hatchets	1.4	and finishing tools>)	2.15	trench knives	2.35
franciscas	1.53	arrows	2.33	arm daggers	2.35
chip axes	1.6	machetes	2.33	dirks	2.35
berdyshes	1.6	darts	2.33	fighting bracelets	2.35
staff weapons	1.75	partisans	2.35	finger hooks	2.35
sword sticks	1.75	spears (weapons)	2.35	finger knives	2.35
harpoons	1.75	leading staffs	2.35	brass knuckles	2.35
bayonets	1.75	bills (staff weapons)	2.35	switchblade knives	2.35
daggers (weapons)	1.75	corsescas	2.35	bolos (weapons)	2.35
fist weapons	1.75	glaives	2.35	bowie knives	2.35
knives (weapons)	1.75	integral bayonets	2.35	<swords by form>	2.35
swords	1.75	knife bayonets	2.35	<swords by function>	2.35
<projectiles with		plug bayonets	2.35	weapons	2.5
nonexplosive propellant>	1.77	socket bayonets	2.35		

One method of reducing noise introduced by RT expansion is to filter on the original term's (sub)hierarchy, in this case *Weapons & Ammunition*. Thus RTs to terms within different sub-hierarchies would not be traversed (or potentially could be penalised). Table 3 shows a set of terms (and their hierarchies) which would be excluded from the previous example in this situation (distances are from Table 2). Note that instances of axes serving both as tools and as weapons (*hatchets*, *machetes*) are now excluded, since due to the mono-hierarchical nature of the AAT they are located within the *Tools&Equipment* hierarchy, and this may sometimes be undesirable.

Table 3. Terms excluded when inter-hierarchical traversals are not allowed. T.&E. stands for Tools & Equipment hierarchy, while I.F. is Information Forms.

Term	Dist.	Sub-hierarchy	Term	Dist.	Sub-hierarchy
axes (tools)	1	T. & E.	<cutting tools>	2	T. & E.
hatchets	1.4	T. & E.	fascies	2	I. F.
chip axes	1.6	T. & E.	Pulaskis	2	T. & E.
adze-hatchets	1.9	T. & E.	<wood-cutting and		
hewing hatchets	1.9	T. & E.	- finishing tools>	2.15	T. & E.
lathing hatchets	1.9	T. & E.	machetes	2.33	T. & E.
shingling hatchets	1.9	T. & E.			

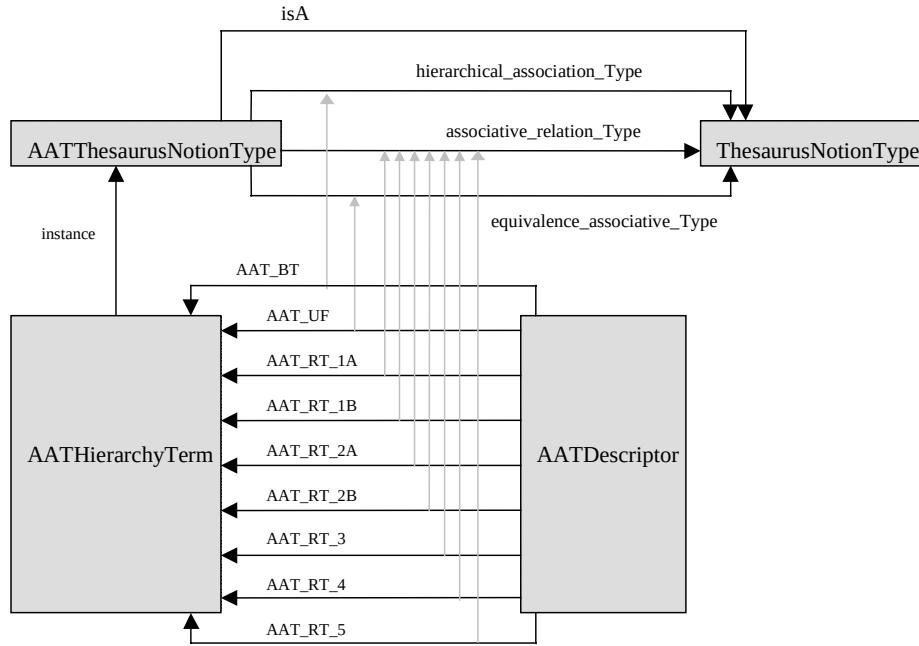


Fig. 3. Specialisation of the associative relationship.

The next scenario explores an alternative approach to filtering based upon selective specialisation of the RT relationship according to retrieval context. This is in keeping with the recommendation of Rada et al [35] that automatic expansion of non-hierarchical relationships be restricted to situations where the type of relationship can be linked with the particular query, and also with Jones' [23] suggestion of using sub-classifications to help distinguish relationships according to strength. The aim is to take advantage of more structured approaches to thesaurus construction where different types of RTs are employed. For example, common subdivisions of RTs include partitive and causal relationships [3]. In some circumstances it may be appropriate to consider all types of associative relationships as a generic RT for retrieval purposes (as in the above scenarios). However, under other contexts it may be desirable to treat RT sub-types differently, permitting some RT traversals but forbidding or penalising (via weighting) others. Thus, heuristics may selectively guide RT expansion, depending on query model and session context. The AAT is particularly suited to investigation of this topic, since its editors followed a systematic, rule-based approach to the design of RT links [30]. The AAT RT editorial manual [1] specifies a set of rules to apply to the relevant hierarchical context and scope notes in order to identify valid RT relationships between terms when building the vocabulary or enhancing it. This includes a set of specialisations of the RT relationships: *1A and 1B*) Alternate hierarchical (BT/NT) relationships (since AAT is not polyhierarchical); *2A and 2B*) Part/Whole relationships; *3*) Several Inter/Intra Facet relationships (eg Agents-Activities and Agents-Materials); *4*) Distinguished From relationship - the scope note evidences a need to distinguish the sense of two

terms; 5) frequently Conjoined terms (eg Cups AND Saucers). We have extended the original SIS AAT schema to specialise the associative relationship. See Figure 3, where (for example) AAT_RT_4 represents the Distinguished From relationship (the 19 AAT_RT_3 subtypes are not displayed separately in interests of space). RTs in our model can optionally be treated as specialised sub-relationships, or as generic RTs via associative_relation_Type.

The editorial rules for creating specific associative relationships are not retained in electronic implementations of the AAT to date. Thus, for this experiment we manually specialised all RT relationships 3 links away from *axes (weapons)* into their corresponding sub-types by following sample extracts of AAT Editorial Related Term Sheets and applying the editorial rules. In the scenario, the distance algorithm was set to filter on the RT subtype, only permitting traversal over the Alternate BT and Alternate NT relationships. Table 4 summarises the differences (terms included and excluded) with the hierarchy filtering approach (Table 3) – all terms of course were present in the unfiltered Table 2. This would correspond to a reasonably strict information request but results retrieved now include terms, such as *machetes*, *hatchets* from the *Tools & Equipment* hierarchy, which were excluded when narrowly filtering on the original hierarchy. For example, an alternate NT relationship exists between *tomahawks* and *hatchets*. Since they are classed as both tools and weapons, *hatchets* might well be regarded as relevant to the scenario.

Table 4. Filtering by RT specialisation.

Terms Included		Terms Excluded	
Term	Dist.	Term	Dist.
hatchets	1.4	axes (tools)	1
adze-hatchets	1.9	chip axes	1.6
hewing hatchets	1.9	<cutting tools>	2
lathing hatchets	1.9	fascies	2
shingling hatchets	1.9		
<wood cutting and -finishing tools>	2.15		
Pulaskis	2.2		
machetes	2.33		

Some reviewers have been critical of the AAT's mono-hierarchical design [38]. The RT specialisations offer an option of treating it as a poly-hierarchical system, for retrieval purposes. It may well be preferable to weight such alternate hierarchical RT relationships identically to BT/NTs, but this is an issue for future investigation.

The AAT Scope Note for *axes (weapons)* reads: “*Cutting weapons consisting basically of a relatively heavy, flat blade fixed to a handle, wielded by either striking or throwing. For axes used for other purposes, typically having narrower blades, use axes (tools).*” Thus the associative relationship between *axes (weapons)* and *axes (tools)* is of subtype *Distinguished From* and is not traversed in the above scenario when filtering on alternate hierarchical RT subtypes. We can see in Table 4 that the term *axes (tools)* and tool-related terms derived solely from this link (*chip axes*, *cutting tools*, etc) are excluded. Under some contexts, such terms might be considered relevant but in a stricter weapons-related scenario they might well be seen as less relevant and can now be suppressed. The point is that this control can be passed to the retrieval system. Other scenarios illustrate the potential for filtering on other types of

RT relationship. For example, an information need relating to *archery and its equipment*, would justify traversal of AAT RT inter-facet subtype *Activity - Equipment Needed or Produced*. This would in turn yield the terms *arrows* and *bows (weapons)*, which could be expanded to terms such as *bolts (arrows)*, *crossbows*, *composite bows*, *longbows*, and *self bows*. The same approach can be applied to scenarios relating to parts or components of an object, using the RT *Whole/Part*, and *Part to Whole* subtypes. Here, a query on *arrows* would yield the terms *nocks* and *<arrow components>* which could be expanded to terms such as *arrowheads*, and *feathers (arrow components)*.

The effect of combining RT and BT/NT expansion, or chains of hierarchical and non-hierarchical relationships warrants some future investigation. Should all possible chains of relationships be considered equally transitive for retrieval purposes? For example, in our scenarios RT-BT traversal chains led to some tenuous links (*<cutting tools>*, *<ceremonial weapons>*). One approach to reducing noise might be to consider penalising certain combinations or vary RT weighting depending on order of relationship traversal, although it is difficult to argue from individual cases. Support for this can be found in the AAT RT editorial manual which stresses a guiding inheritance principle when identifying RT relationships: RT links from an initial terms must apply to all NTs of the target term. RT-BT chains could be seen as less valid and RT-NT chains as more valid from consideration of the inheritance principle – however the topic needs further investigation.

5 Conclusions

It may be impractical to expect non-specialist users to manually browse very large thesauri (for example, there are 1792 terms in the AAT's *Tools&Equipment* hierarchy). Semantic distance measures operating over thesaurus relationships can underpin interactive and automatic query expansion techniques by ranking candidate query terms or results. Results are presented in this paper from novel approaches to semantic distance measures for associative relationships and geographical thesauri. Online gazetteers and geographical thesauri may not contain co-ordinate data for all places and regions or, if they do, associate place names with a limited spatial footprint (centroid or minimum bounding rectangle). In such situations, the ability to rank places within a vicinity according to hierarchical (or other) relationships in a spatial terminology system can be useful. In contexts where administrative boundaries are highly relevant, distance measures could combine quantitative and qualitative spatial relationships. Related work has noted the potential of RTs in thesaurus search aids but the problem of increased noise in result sets has been emphasised. Experimental scenarios (Section 4) exploring different factors relating to incorporation of RTs in semantic distance measures demonstrate the potential for filtering on the context of the RT link in faceted thesauri and on subtypes of RT relationships. Specialising RTs allows the possibility of dynamically linking RT type to query context and, in cases like the AAT, treating alternate hierarchical RT relationships more flexibly for retrieval purposes. Thus RT subtypes could be selectively filtered in or out of distance measures, depending on cues derived from an expression of information need or from information elicited by a query editor. In practice, it is likely that a combination of filtering heuristics will be useful. An ability for retrieval systems to optionally

specialise RTs or to treat them as generic would retain the advantages of the standard core set of thesaurus relationships for interoperability purposes – some thesauri or terminology systems will only contain the core relationships. However, the ability to deal with a richer semantics of RT sub-relationships would allow more flexibility in retrieval where it was possible. Note that it is also possible to specialise hierarchical relationships².

There are implications for thesaurus developers and implementers. A systematic approach to RT application in thesaurus design, as in the AAT, has potential for retrieval systems. Information (eg of relationship subclasses) used in thesaurus design should be retained in data models and database design for later use in retrieval algorithms. In future work, we intend to build on the underlying semantic distance measures and explore how best to incorporate thesaurus semantic distance controls in the user interface. The issue of RT specialisations expressing thesaurus inter-facet links and the retrieval implications is a promising area, which converges with work on broader ontological conceptualisations attempting to more formally define the roles played by entities in the schema.

Acknowledgements

We would like to thank the Getty Information Institute for provision of their vocabularies and in particular Alison Chipman for information on Related Terms; Diana Murray and the Royal Commission on the Ancient and Historical Monuments of Scotland for provision of their dataset; and Martin Doerr and Christos Georgis from the FORTH Institute of Computer Science for assistance with the SIS.

References

1. AAT 1995. The AAT Editorial Manual: Related terms. User Friendly, 2(3-4), 6-15. Getty Art History Information Program.
2. AAT 2000. http://shiva.pub.getty.edu/aat_browser.
3. Aitchison J., Gilchrist A. 1987. Thesaurus construction: a practical manual. ASLIB: London.
4. Alani H., Jones C., Tudhope D. *in press*. Voronoi-based region approximation for geographical information retrieval with online gazetteers. *Internat. Journal of Geographic Information Systems*.
5. Amann B., Fundulaki I. 1999. Integrating ontologies and thesauri to build RDF schemas. *Proc. 3rd European Conference on Digital Libraries (ECDL'99)*, (S. Abiteboul and A. Vercoustre eds.) *Lecture Notes in Computer Science 1696*, Springer-Verlag: Berlin, 234-253.
6. Beaulieu M. 1997. Experiments on interfaces to support query expansion. *Journal of Documentation*, 53(1), 8-19.
7. Bosman F., Bruza P., van der Weide T., Weusten L. 1998. Documentation, cataloguing, and query by navigation: a practical and sound approach. *Proc. 2nd European Conference on Digital Libraries (ECDL'98)*, (C. Nikolaou and C. Stephanidis eds.) *Lecture Notes in Computer Science 1513*, Springer-Verlag: Berlin, 459-478.
8. Brooks T. 1997. The relevance aura of bibliographic records. *Information Processing and Management*, 33(1), 69-80.
9. Chen H., Dhar V. 1991. Cognitive process as a basis for intelligent retrieval systems design. *Information Processing and Management*, 27(5), 405-432.

² For examples of RDF representations of both a core set of thesaurus relationships and a more complex set of relationships, see <http://www.desire.org/results/discovery/rdfthesschema.html> (Cross, Brickley & Koch)

10. Chen H., Ng T., Martinez J., Schatz B. 1997. A concept space approach to addressing the vocabulary problem in scientific information retrieval: an experiment on the Worm Community System. *Journal of the American Society for Information Science*, 48(1), 17-31.
11. Cohen, P. R. and R. Kjeldsen (1987). Information Retrieval by Constrained Spreading Activation in Semantic Networks. *Information Processing & Management* 23(4): 255-268.
12. Constantopolous P., Doerr M. 1993. The Semantic Index System - A brief presentation. Institute of Computer Science Technical Report. FORTH-Hellas, GR-71110 Heraklion, Crete.
13. Croft W., Lucia T., Cringean J., Willett P. 1989. Retrieving documents by plausible inference: an experimental study. *Information Processing and Management*, 25(6), 599-614.
14. Cunliffe D., Taylor C., Tudhope D. 1997. Query-based navigation in semantically indexed hypermedia. *Proc. 8th ACM Conference on Hypertext*, 87-95.
15. Doerr M., Fundulaki I. 1998. SIS-TMS: A thesaurus management system for distributed digital collections. *Proc. 2nd European Conference on Digital Libraries (ECDL'98)*, (C. Nikolaou and C. Stephanidis eds.) *Lecture Notes in Computer Science* 1513, Springer-Verlag: Berlin, 215-234.
16. Fidel R. 1991. Searchers' selection of search keys (I-III), *Journal of American Society for Information Science*, 42(7), 490-527.
17. Frew J., Freeston M., Freitas N., Hill L., Janee G., Lovette K., Nideffer R., Smith T., Zheng Q. 1998. The Alexandria Digital Library Architecture. *Proc. 2nd European Conference on Digital Libraries (ECDL'98)*, (C. Nikolaou and C. Stephanidis eds.) *Lecture Notes in Computer Science* 1513, Springer-Verlag: Berlin, 61-73.
18. Guarino N. 1995. Ontologies and knowledge bases: towards a terminological clarification. In: *Towards very large knowledge bases: knowledge building and knowledge sharing*, 25-32. IOS Press.
19. Harper Collins, 2000, Bartholomew. <http://www.bartholomewmaps.com>
20. Harpring P. 1997. The limits of the world: Theoretical and practical issues in the construction of the Getty Thesaurus of Geographic Names. *Proc. 4th International Conference on Hypermedia and Interactivity in Museums (ICHIM'97)*, 237-251, *Archives and Museum Informatics*.
21. Harpring P. 1999. How forcible are the right words: overview of applications and interfaces incorporating the Getty vocabularies. *Proc. Museums and the Web 1999. Archives and Museum Informatics*. <http://www.archimuse.com/mw99/papers/harpring/harpring.html>
22. Jones C. 1997. Geographic Interfaces to Museum Collections. *Proc. 4th International Conference on Hypermedia and Interactivity in Museums (ICHIM'97)*, 226-236, *Archives and Museum Informatics*.
23. Jones, S. 1993. A Thesaurus Data Model for an Intelligent Retrieval System. *Journal of Information Science* 19: 167-178.
24. Jones S., Gatford M., Robertson S., Hancock-Beaulieu M., Secker J., Walker S. 1995. Interactive Thesaurus Navigation: Intelligence Rules OK?, *Journal of the American Society for Information Science*, 46(1), 52-59.
25. Kim Y., Kim J. 1990. A model of knowledge based information retrieval with hierarchical concept graph. *Journal of Documentation*, 46(2), 113-136.
26. Kristensen J. 1993. Expanding end-users' query statements for free text searching with a search-aid thesaurus. *Information Processing and Management*, 29(6), 733-744.
27. Lee J., Kim H., Lee Y. 1993. Information retrieval based on conceptual distance in ISA hierarchies. *Journal of Documentation*, 49(2), 113-136.
28. McMath C. F., Tamaru R. S., Rada R. 1989. A graphical thesaurus-based information retrieval system, *International Journal of Man-Machine Studies*, 31(2), 121-147.
29. Michard A., Pham-Dac G. 1998. Description of Collections and Encyclopaedias on the Web using XML. *Archives and Museum Informatics*, 12(1), 39-79.
30. Molholt P. 1996. Standardization of inter-concept links and their usage. *Proc. 4th International ISKO Conference, Advances in Knowledge Organisation* (5), 65-71.
31. Murray D. 1997. GIS in RCAHMS. *MDA Information* 2(3): 35-38.
32. Paice C. 1991. A thesaural model of information retrieval. *Information Processing and Management*, 27(5), 433-447.
33. Pollitt A. 1997. Interactive information retrieval based on faceted classification using views. *Proc. 6th International Study Conference on Classification*, London.
34. Rada R., Mili H., Bicknell E., Blettner M. (1989). Development and Application of a Metric on Semantic Nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19(1), 17-30.
35. Rada R., Barlow J., Potharst J., Zanzara P., Bijstra D. 1991. Document ranking using an enriched thesaurus. *Journal of Documentation*, 47(3), 240-253.
36. Richardson R., Smeaton A., Murphy J. 1994. Using Wordnet for conceptual distance measurement, *Proc. 16th Research Colloquium of BCS IR Specialist Group*, 100-123.

37. Smeaton A., & Quigley I. 1996. Experiments on Using Semantic Distances Between Words in Image Caption Retrieval, Proc. 19th ACM SIGIR Conference, 174-180.
38. Soergel. D 1995. The Art and Architecture Thesaurus (AAT): a critical appraisal. Visual Resources, 10(4), 369-400.
39. Spanoudakis G., Constantopoulos P. 1994. Similarity for analogical software reuse: a computational model. Proc. 11th European Conference on AI (ECAI'94), 18-22. Wiley.
40. Spanoudakis G., Constantopoulos P. 1996. Elaborating analogies from conceptual models. International Journal of Intelligent Systems. 11, 917-974.
41. Tudhope D., Taylor C. 1997. Navigation via Similarity: automatic linking based on semantic closeness. Information Processing and Management, 33(2), 233-242.
42. Tudhope D., Cunliffe D. 1999. Semantic index hypermedia: linking information disciplines. ACM Computing Surveys, Symposium on Hypertext and Hypermedia. *in press*.

Experiments on the Use of Feature Selection and Negative Evidence in Automated Text Categorization

Luigi Galavotti¹, Fabrizio Sebastiani², and Maria Simi³

¹ AUTON S.R.L.

Via Jacopo Nardi, 2 – 50132 Firenze, Italy

galavott@tin.it

² Istituto di Elaborazione dell'Informazione – Consiglio Nazionale delle Ricerche

56100 Pisa, Italy

fabrizio@iei.pi.cnr.it

³ Dipartimento di Informatica – Università di Pisa

56125 Pisa, Italy

simi@di.unipi.it

Abstract. We tackle two different problems of *text categorization* (TC), namely feature selection and classifier induction. *Feature selection* (FS) refers to the activity of selecting, from the set of r distinct features (i.e. words) occurring in the collection, the subset of $r' \ll r$ features that are most useful for compactly representing the meaning of the documents. We propose a novel FS technique, based on a simplified variant of the χ^2 statistics. *Classifier induction* refers instead to the problem of automatically building a text classifier by learning from a set of documents pre-classified under the categories of interest. We propose a novel variant, based on the exploitation of negative evidence, of the well-known k -NN method. We report the results of systematic experimentation of these two methods performed on the standard REUTERS-21578 benchmark.

1 Introduction

Text categorization (TC) denotes the activity of automatically building, by means of machine learning techniques, automatic text classifiers, i.e. systems capable of labelling natural language texts with thematic categories from a predefined set $C = \{c_1, \dots, c_m\}$ (see e.g. [6]). In general, this is actually achieved by building m independent classifiers, each capable of deciding whether a given document d_j should or should not be classified under category c_i , for $i \in \{1, \dots, m\}$ ¹. This process requires the availability of a corpus $Co = \{d'_1, \dots, d'_s\}$ of preclassified documents, i.e. documents such that for all $i \in \{1, \dots, m\}$ and for all

¹ We here make the assumption that a document d_j can belong to zero, one or many of the categories in C ; this assumption is verified in the REUTERS-21578 benchmark we use for our experiments. All the techniques we discuss here can be straightforwardly adapted to the other case in which each document belongs to exactly one category.

$j \in \{1, \dots, s\}$ it is known whether $d'_j \in c_i$ or not. A general inductive process (called the *learner*) automatically builds a classifier for category c_i by learning the characteristics of c_i from a *training set* $Tr = \{d'_1, \dots, d'_g\} \subset Co$ of documents. Once a classifier has been built, its effectiveness (i.e. its capability to take the right categorization decisions) may be tested by applying it to the *test set* $Te = \{d'_{g+1}, \dots, d'_s\} = Co - Tr$ and checking the degree of correspondence between the decisions of the automatic classifier and those encoded in the corpus.

Two key steps in the construction of a text classifier are document indexing and classifier induction. *Document indexing* refers to the task of automatically constructing internal representations of the documents that (i) be amenable to interpretation by the classifier induction algorithm (and by the text classifier itself, once this has been built), and (ii) compactly capture the meaning of the documents. Usually, a text document is represented as a vector of weights $d_j = \langle w_{1j}, \dots, w_{rj} \rangle$, where r is the number of features (i.e. words) that occur at least once in at least one document of Co , and $0 \leq w_{kj} \leq 1$ represents, loosely speaking, how much feature t_k contributes to the semantics of document d_j . Many classifier induction methods are computationally hard, and their computational cost is a function of r . It is thus of key importance to be able to work with vectors shorter than r , which is usually a number in the tens of thousands or more. For this, *feature selection* (FS) techniques are used to select, from the original set of r features, a subset of $r' \ll r$ features that are most useful for compactly representing the meaning of the documents. In this work we propose a novel technique for FS based on a simplified variant of the χ^2 statistics; we call this technique *simplified χ^2* . The key issues of FS and our simplified χ^2 method are introduced in Section 2, while the results of its extensive experimentation on REUTERS-21578, the standard benchmark of TC research, are described in Section 4.1.

Classifier induction refers instead to the inductive construction of a text classifier from a training set of documents that have already undergone indexing and FS. We propose a novel classifier induction technique based on a variant of k -NN, a popular instance-based method. After introducing instance-based methods in Section 3, in Section 3.1 we describe our modified version of k -NN, based on the exploitation of negative evidence. The results of its experimentation on REUTERS-21578 are described in Section 4.2. Section 5 concludes.

2 Issues in feature selection

Given a fixed $r' \ll r$, the aim of FS is to select, from the original set of r features that occur at least once in at least one document in Co , the r' features that, when used for document indexing, yield the best categorization effectiveness. The value $(1 - \frac{r'}{r})$ is called the *aggressivity* of the selection; the higher this value, the smaller the set resulting from FS, and the higher the computational benefits. On the other hand, a high aggressivity may curtail the ability of the classifier to correctly “understand” the meaning of a document, since information that in

principle may contribute to specify document meaning is removed. Therefore, deciding on the best level of aggressivity usually requires some experimentation.

A widely used approach to FS is the so-called *filtering* approach, which consists in selecting the $r' \ll r$ features that score highest according to a function that measures the “importance” of the feature for the categorization task. In a thorough comparative experiment, performed across different classifier induction methods and different benchmarks, Yang and Pedersen [10] have shown

$$\chi^2(t_k, c_i) = \frac{g[P(t_k, c_i)P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i)P(\bar{t}_k, c_i)]^2}{P(t_k)P(\bar{t}_k)P(c_i)P(\bar{c}_i)} \quad (1)$$

to be one of the most effective functions for the filtering method, allowing aggressivity levels in the range [.90,.99] with no loss (or even with a small increase) of effectiveness. This contributes to explain the popularity of χ^2 as a FS technique in TC (see [6, Section 5]).

In Equation 1 and in those that follow, g indicates the cardinality of the training set, and probabilities are interpreted on an event space of documents (e.g. $P(\bar{t}_k, c_i)$ indicates the probability that, for a random document x , feature t_k does not occur in x and x belongs to category c_i), and are estimated by counting occurrences in the training set. Also, every function $f(t_k, c_i)$ discussed in this section evaluates the feature with respect to a specific category c_i ; in order to assess the value of a feature t_k in a “global”, category-independent sense, either the weighted average $f_{avg}(t_k) = \sum_{i=1}^m f(t_k, c_i)P(c_i)$ or the maximum $f_{max}(t_k) = \max_{i=1}^m f(t_k, c_i)$ of its category-specific values are usually computed.

In the experimental sciences χ^2 is used to measure how the results of an observation differ from those expected according to an initial hypothesis. In our application the initial hypothesis is that t_k and c_i are independent, and the truth of this hypothesis is “observed” on the training set. The features t_k with the lowest value for $\chi^2(t_k, c_i)$ are thus the most independent from c_i ; as we are interested in those features which are not, we select those features for which $\chi^2(t_k, c_i)$ is highest.

However, Ng et al. [4] have observed that some aspects of χ^2 clash with the intuitions that underlie FS. In particular, they observe that the power of 2 at the numerator has the effect of equating the roles of the probabilities that indicate a positive correlation between t_k and c_i (i.e. $P(t_k, c_i)$ and $P(\bar{t}_k, \bar{c}_i)$) and those that indicate a negative correlation (i.e. $P(t_k, \bar{c}_i)$ and $P(\bar{t}_k, c_i)$). The function

$$CC(t_k, c_i) = \frac{\sqrt{g}[P(t_k, c_i)P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i)P(\bar{t}_k, c_i)]}{P(t_k)P(\bar{t}_k)P(c_i)P(\bar{c}_i)} \quad (2)$$

they propose, being the square root of $\chi^2(t_k, c_i)$, emphasizes thus the former and de-emphasizes the latter. The experimental results by Ng et al. [4] show a superiority of $CC(t_k, c_i)$ over $\chi^2(t_k, c_i)$.

In this work we go a further step in this direction, by observing that in $CC(t_k, c_i)$, and *a fortiori* in $\chi^2(t_k, c_i)$:

- The $\sqrt{g}$ factor at the numerator is redundant, since it is equal for all pairs (t_k, c_i) . This factor can thus be removed.

- The presence of $\frac{j}{P(t_k)P(\bar{t}_k)}$ at the denominator emphasizes very rare features, since for these features it has very low values. By showing that document frequency is a very effective FS technique, [10] has shown that very rare features are the least effective in TC. This factor should thus be removed.
- The presence of $\frac{j}{P(c_i)P(\bar{c}_i)}$ at the denominator emphasizes very rare categories, since for these categories this factor has very low values. Emphasizing very rare categories is counterintuitive, since this tends to depress microaveraged effectiveness (see Section 4), which is now considered the correct way to measure effectiveness in most applications [6, Section 8]. This factor should thus be removed.

Removing these three factors from $CC(t_k, c_i)$ yields

$$s\chi^2(t_k, c_i) = P(t_k, c_i)P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i)P(\bar{t}_k, c_i) \quad (3)$$

In Section 4 we discuss the experiments we have performed with $s\chi^2(t_k, c_i)$ on the REUTERS-21578 benchmark.

3 Issues in instance-based classifier induction

One of the most popular paradigms for the inductive construction of a classifier is the *instance-based* approach, which is well exemplified by the k -NN (for “ k nearest neighbors”) algorithm used e.g. by Yang [7]. For deciding whether d_j should be classified under c_i , k -NN selects the k training documents most similar to d_j . Those documents d'_z that belong to c_i are seen as carrying evidence towards the fact that d_j also belongs to c_i , and the amount of this evidence is proportional to the similarity between d'_z and d_j . Classifying a document with k -NN thus means computing

$$CSV_i(d_j) = \sum_{d'_z \in Tr_k(d_j)} RSV(d_j, d'_z) \cdot v_{iz} \quad (4)$$

where

- $CSV_i(d_j)$ (the *categorization status value* of document d_j for category c_i) measures the computed evidence that d_j belongs to c_i ;
- $RSV(d_j, d'_z)$ (the *retrieval status value* of document d'_z with respect to document d_j) represents a measure of semantic relatedness between d_j and d'_z ;
- $Tr_k(d_j)$ is the set of the k training documents d'_z with the highest $RSV(d_j, d'_z)$;
- the value of v_{iz} is given by

$$v_{iz} = \begin{cases} 1 & \text{if } d'_z \text{ is a positive instance of } c_i \\ 0 & \text{if } d'_z \text{ is a negative instance of } c_i \end{cases}$$

The threshold k , indicating how many top-ranked training documents have to be considered for computing $CSV_i(d_j)$, is usually determined experimentally on a validation set; Yang [7, 8] has found $30 \leq k \leq 45$ to yield the best effectiveness.

Usually, the construction of a classifier, instance-based or not, also involves the determination of a threshold τ_i such that $CSV_i(d_j) \geq \tau_i$ may be viewed as an indication to file d_j under c_i and $CSV_i(d_j) < \tau_i$ may be viewed as an indication not to file d_j under c_i . For determining this threshold we have used the *proportional thresholding* method, as in our experiments this has proven superior to *CSV thresholding* (see [6, Section 7]).

3.1 Using negative evidence in instance-based classification

The basic philosophy that underlies k -NN and all the instance-based algorithms used in the TC literature may be summarized by the following principle:

Principle 1 *If a training document d'_z similar to the test document d_j is a positive instance of category c_i , then use this fact as evidence towards the fact that d_j belongs to c_i . Else, if d'_z is a negative instance of c_i , do nothing.*

The first part of this principle is no doubt intuitive. Suppose d_j is a news article about Reinhold Messner's ascent of Mt. Annapurna, and d'_z is a very similar document, e.g. a news account of Anatoli Bukreev's expedition to Mt. Everest. It is quite intuitive that if d'_z is a positive instance of category REUTRUS, this information should carry evidence towards the fact that d_j too is a positive instance of REUTRUS. But the same example shows, in our opinion, that the second part of this principle is unintuitive, as the information that d'_z is a negative instance of category 2157U8S should not be discarded, but should carry evidence towards the fact that d_j too is a negative instance of 2157U8S.

In this work, we thus propose a variant of the k -NN approach in which *negative evidence* (i.e. evidence provided by negative training instances) is not discarded. This may be viewed as descending from a new principle:

Principle 2 *If a training document d'_z similar to the test document d_j is a positive instance of category c_i , then use this fact as evidence towards the fact that d_j belongs to c_i . Else, if d'_z is a negative instance of c_i , then use this fact as evidence towards the fact that d_j does not belong to c_i .*

Mathematically, this comes down to using

$$v_{iz} = \begin{cases} 1 & \text{if } d'_z \text{ is a positive instance of } c_i \\ -1 & \text{if } d'_z \text{ is a negative instance of } c_i \end{cases}$$

in Equation 4. We call the method deriving from this modification $k\text{-NN}_{neg}^1$ (this actually means $k\text{-NN}_{neg}^p$ for $p = 1$; the meaning of the p parameter will become clear later). This method brings instance-based learning closer to most other classifier induction methods, in which negative training instances play a fundamental role in the individuation of a “best” decision surface (i.e. classifier) that separates positive from negative instances. Even methods like Rocchio (see [6, Section 6]), in which negative instances had traditionally been either discarded or at best de-emphasized, have recently been shown to receive a performance boost by an appropriate use of negative instances [5].

4 Experimental results

In our experiments we have used the “REUTERS-21578, Distribution 1.0” corpus, as it is currently the most widely used benchmark in TC research². REUTERS-21578 consists of a set of 12,902 news stories, partitioned (according to the “ModApté” split we have adopted) into a training set of 9,603 documents and a test set of 3,299 documents. The documents are labelled by 118 categories; the average number of categories per document is 1.08, ranging from a minimum of 0 to a maximum of 16. The number of positive instances per category ranges from a minimum of 1 to a maximum of 3964.

We have run our experiments on the set of 115 categories with at least 1 training instance, rather than on other subsets of it. The full set of 115 categories is “harder”, since it includes categories with very few positive instances for which inducing reliable classifiers is obviously a haphazard task. This explains the smaller effectiveness values we have obtained with respect to experiments carried out by other researchers with exactly the same methods but on reduced REUTERS-21578 category sets (e.g. the experiments reported in [9] with standard k -NN). In all the experiments discussed in this section, stop words have been removed using the stop list provided in [3, pages 117–118]. Punctuation has been removed and letters have been converted to lowercase; no stemming and number removal have been performed. Term weighting has been done by means of the standard “ltc” variant of the $tf * idf$ function. Classification effectiveness has been measured in terms of the classic IR notions of precision (Pr) and recall (Re) adapted to the case of document categorization. We have evaluated *microaveraged* precision and recall, since it is almost universally preferred to *macroaveraging* [6, Section 8]. As a measure of effectiveness that combines the contributions of both Pr and Re , we have used the well-known function $F_1 = \frac{2 \cdot Pr \cdot Re}{Pr + Re}$. See the full paper for more details on the experiments.

4.1 Feature selection experiments

We have performed our FS experiments first with the standard k -NN classifier of Section 3 (with $k = 30$), and subsequently with a Rocchio classifier we have implemented following [1] (the Rocchio parameters were set to $\beta = 16$ and $\gamma = 4$; see [1, 5] for a full discussion of the Rocchio method). In these experiments we have compared two baseline FS functions, i.e. $\#_{avg}(t_k) = \frac{1}{k} \sum_{i=1}^k \#(t_k, c_i)P(c_i)$ and $\chi_{max}^2(t_k) = \max_{i=1}^m \chi^2(t_k, c_i)$, to two variants of our $s\chi^2(t_k)$ function, i.e. $s\chi_{max}^2(t_k) = \max_{i=1}^m s\chi^2(t_k, c_i)$ and $s\chi_{avg}^2(t_k) = \frac{1}{k} \sum_{i=1}^k s\chi^2(t_k, c_i)P(c_i)$. As a baseline, we have chosen $\chi_{max}^2(t_k)$ and not $\chi_{avg}^2(t_k)$ because the former is known to perform substantially better than the latter [10]. Table 1 lists the microaveraged F_1 values for k -NN and Rocchio with different FS techniques at different aggressivity levels. A few conclusions may be drawn from these results:

² The Reuters-21578 corpus may be freely downloaded for experimentation purposes from <http://www.research.att.com/~lewis/reuters21578.html>

Reduction level	k -NN				Rocchio			
	$\#(t_k)$	$\chi_{max}^2(t_k)$	$s\chi_{max}^2(t_k)$	$s\chi_{avg}^2(t_k)$	$\#(t_k)$	$\chi_{max}^2(t_k)$	$s\chi_{max}^2(t_k)$	$s\chi_{avg}^2(t_k)$
99.9	—	—	—	—	.458	.391	.494	—
99.5	—	—	—	—	.624	.479	.657	—
99.0	.671	.648	.697	.501	.656	.652	.692	—
98.0	.703	.720	.734	.554	.691	.710	.736	—
96.0	.721	.766	.729	.577	.737	.733	.748	—
94.0	.731	.766	.728	.596	—	—	—	—
92.0	.729	.772	.732	.607	—	—	—	—
90.0	.734	.775	.732	.620	—	—	—	—
85.0	.735	.767	.726	.640	—	—	—	—
80.0	.734	.757	.730	.658	—	—	—	—
70.0	.734	.748	.730	.682	—	—	—	—
60.0	.732	.741	.733	.691	—	—	—	—
50.0	.733	.735	.734	.701	—	—	—	—
40.0	.733	.735	.731	.716	—	—	—	—
30.0	.731	.732	.730	.721	—	—	—	—
20.0	.731	.732	.730	.727	—	—	—	—
10.0	.730	.730	.730	.730	—	—	—	—
00.0	.730	.730	.730	.730	—	—	—	—

Table 1. Microaveraged F_1 values for k -NN ($k = 30$) and Rocchio ($\alpha = 16$ and $\beta = 4$).

- on the k -NN tests we performed first, $s\chi_{avg}^2(t_k)$ proved largely inferior to $s\chi_{max}^2(t_k)$ (and to all other FS functions tested). This is reminiscent of Yang and Pedersen’s [10] result, who showed that $\chi_{avg}^2(t_k)$ is outperformed by $\chi_{max}^2(t_k)$. As a consequence, due to time constraints we have abandoned $s\chi_{avg}^2(t_k)$ without further testing it on Rocchio;
- on the k -NN tests, $s\chi_{max}^2(t_k)$ is definitely inferior to $\chi_{max}^2(t_k)$ and comparable to $\#_{avg}(t_k)$ up to levels of reduction around .95, but becomes largely superior for aggressivity levels higher than that;
- following this observation, we have run Rocchio tests with extreme (from .960 up to .999) aggressivity levels, and observed that in these conditions $s\chi_{max}^2(t_k)$ outperforms both $\chi_{max}^2(t_k)$ and $\#_{avg}(t_k)$ by a wide margin.

The conclusion we may draw from these experiments is that $s\chi_{max}^2(t_k)$ is a superior alternative to both $\chi_{max}^2(t_k)$ and $\#_{avg}(t_k)$ when very aggressive FS is necessary. Besides, it is important to remark that $s\chi_{max}^2(t_k)$ is much easier to compute than $\chi_{max}^2(t_k)$. Altogether, these facts indicate that $s\chi_{max}^2(t_k)$ may be a very good choice in the context of learning algorithms that do not scale well to high dimensionalities of the feature space, such as neural networks, or in the application to TC tasks characterized by very high dimensionalities.

4.2 Classifier induction experiments

We have performed our classifier induction experiments by comparing a standard k -NN algorithm with our modified k -NN_{neg}¹ method, at different values of k . For

FS we have chosen $\chi_{max}^2(t_k)$ with .90 aggressivity since this had yielded the highest effectiveness ($F_1 = .775$) in the experiments of Section 4.1. The results of this experimentation are reported in the 1st and 2nd columns of Table 2.

k	k -NN			k -NN $_{neg}^1$			k -NN $_{neg}^2$			k -NN $_{neg}^3$		
	Re	Pr	F_1	Re	Pr	F_1	Re	Pr	F_1	Re	Pr	F_1
05	.711	.823	.763	.667	.821	.737	.709	.825	.764	.711	.823	.764
10	.718	.830	.770	.671	.918	.775	.720	.837	.774	.722	.834	.774
20	.722	.833	.774	.663	.930	.774	.725	.841	.780	.725	.836	.778
30	.714	.846	.775	.647	.931	.763	.722	.861	.787	.721	.854	.782
40	.722	.834	.774	.638	.934	.765	.731	.854	.786	.730	.841	.781
50	.724	.836	.776	.628	.938	.752	.730	.854	.786	.730	.843	.782
60	.724	.835	.776	.617	.940	.745	.730	.850	.785	.730	.842	.782
70	.722	.833	.774	.611	.945	.742	.731	.851	.786	.730	.842	.782

Table 2. Experimental comparison between k -NN and k -NN $_{neg}^p$ for different values of k and p , performed with χ_{max}^2 FS and aggressivity .90, and evaluated by microaveraging.

A few observations may be made:

1. Bringing to bear negative evidence in the learning process has not brought about the performance improvement we had expected. In fact, the highest performance obtained for k -NN $_{neg}^1$ (.775) is practically the same as that obtained for k -NN (.776).
2. The performance of k -NN $_{neg}^1$ peaks at substantially lower values of k than for k -NN (10 vs. 50), i.e. much fewer training documents similar to the test document need to be examined for k -NN $_{neg}^1$ than for k -NN.
3. k -NN $_{neg}^1$ is a little less robust than k -NN with respect to the choice of k . In fact, for k -NN $_{neg}^1$ effectiveness degrades somehow for values of k higher than 10, while for k -NN it is hardly influenced by the value of k .

Observation 1 seems to suggest that negative evidence is not detrimental to the learning process, while Observation 2 indicates that, under certain conditions, it may actually be valuable. Instead, we interpret Observation 3 as indicating that negative evidence brought by training documents that are not very similar to the test document may be detrimental.

This is indeed intuitive. Suppose d_j is our news article about Reinhold Messner's ascent of Mt. Annapurna, and d'_z is a critical review of a Picasso exhibition. Should the information that d'_z is a negative instance of category c_i carry any evidence at all towards the fact that d_j too is a negative instance of c_i ? Hardly so, given the wide semantic distance that separates the two texts. While very dissimilar documents have not much influence in k -NN, since positive instances are usually far less than negative ones, they do in k -NN $_{neg}^1$, since each of the k most similar documents, however semantically distant, brings a little weight to the final sum of which the CSV consists.

A similar observation lies at the heart of the use of “query zoning” techniques in the context of Rocchio classifiers [5]; here, the idea is that in learning a concept, the most interesting negative instances of this concept are “the least negative ones” (i.e. the negative instances most similar to the positive ones), in that they are more difficult to separate from the positive instances. Similarly, support vector machine classifiers [2] are induced by using just the negative instances closest to the decision surfaces (i.e. the so-called *negative support vectors*), while completely forgetting about the others.

A possible way to exploit this observation is switching to CSV functions that downplay the influence of the similarity value in the case of widely dissimilar documents; a possible class of such functions is

$$CSV_i(d_j) = \frac{RSV(d_j, d'_z)^p \cdot v_{iz}}{d'_z \in Tr_k(d_j)} \quad (5)$$

in which the larger the value of the p parameter is, the more the influence of the similarity value is downplayed in the case of widely dissimilar documents. We call this method $k\text{-NN}_{neg}^p$. We have run an initial experiment, whose results are reported in the third and fourth column of Table 2 and which has confirmed our intuition: $k\text{-NN}_{neg}^2$ systematically outperforms not only $k\text{-NN}_{neg}^1$ but also standard $k\text{-NN}$. The $k\text{-NN}_{neg}^2$ method peaks for a higher value of k than $k\text{-NN}_{neg}^1$ and is remarkably more stable for higher values of k . This seemingly suggests that negative evidence provided by very dissimilar documents is indeed useful, provided its importance is de-emphasized. Instead, $k\text{-NN}_{neg}^3$ slightly underperforms $k\text{-NN}_{neg}^2$, showing that the level of de-emphasization must be chosen carefully.

5 Conclusion and further research

In this paper we have discussed two novel techniques for TC: $s\chi^2$, a FS technique based on a simplified version of χ^2 , and $k\text{-NN}_{neg}^p$, a classifier learning method consisting of a variant, based on the exploitation of negative evidence, of the popular $k\text{-NN}$ instance-based method.

Concerning the former method, in experiments performed on REUTERS-21578 simplified χ^2 has systematically outperformed χ^2 , one of the most popular FS techniques, at very aggressive levels of reduction, and has done so by a wide margin. This fact, together with its low computational cost, make simplified χ^2 a very attractive method in those applications which demand radical reductions in the dimensionality of the feature space.

Concerning $k\text{-NN}_{neg}^p$, our hypothesis that evidence contributed by negative instances could provide an effectiveness boost for the TC task has been only partially confirmed by the experiments. In fact, our $k\text{-NN}_{neg}^1$ method has performed as well as the original $k\text{-NN}$ but no better than it, and has furthermore shown to be more sensitive to the choice of k than the standard version. However, we have shown that by appropriately de-emphasizing the importance of very dissimilar training instances this method consistently outperforms standard $k\text{-NN}$. Given the prominent role played by $k\text{-NN}$ in the TC literature, and given the simple

modification that moving from k -NN to k -NN $_{neg}^p$ requires, we think this is an interesting result.

References

- [1] D. J. Ittner, D. D. Lewis, and D. D. Ahn. Text categorization of low quality images. In *Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval*, pages 301–315, Las Vegas, US, 1995.
- [2] T. Joachims. Text categorization with support vector machines: learning with many relevant features. In C. Nédellec and C. Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398 in Lecture Notes in Computer Science, pages 137–142, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.
- [3] D. D. Lewis. *Representation and learning in information retrieval*. PhD thesis, Department of Computer Science, University of Massachusetts, Amherst, US, 1992.
- [4] H. T. Ng, W. B. Goh, and K. L. Low. Feature selection, perceptron learning, and a usability case study for text categorization. In N. J. Belkin, A. D. Narasimhalu, and P. Willett, editors, *Proceedings of SIGIR-97, 20th ACM International Conference on Research and Development in Information Retrieval*, pages 67–73, Philadelphia, US, 1997. ACM Press, New York, US.
- [5] R. E. Schapire, Y. Singer, and A. Singhal. Boosting and Rocchio applied to text filtering. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, pages 215–223, Melbourne, AU, 1998. ACM Press, New York, US.
- [6] F. Sebastiani. Machine learning in automated text categorisation: a survey. Technical Report IEI-B4-31-1999, Istituto di Elaborazione dell’Informazione, Consiglio Nazionale delle Ricerche, Pisa, IT, 1999.
- [7] Y. Yang. Expert network: effective and efficient learning from human decisions in text categorisation and retrieval. In W. B. Croft and C. J. van Rijsbergen, editors, *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval*, pages 13–22, Dublin, IE, 1994. Springer Verlag, Heidelberg, DE.
- [8] Y. Yang. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1-2):69–90, 1999.
- [9] Y. Yang and X. Liu. A re-examination of text categorization methods. In M. A. Hearst, F. Gey, and R. Tong, editors, *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval*, pages 42–49, Berkeley, US, 1999. ACM Press, New York, US.
- [10] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In D. H. Fisher, editor, *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 412–420, Nashville, US, 1997. Morgan Kaufmann Publishers, San Francisco, US.

The Benefits of Displaying Additional Internal Document Information on Textual Database Search Result Lists

Offer Drori

The Hebrew University of Jerusalem
Shaam - Information Systems
P. O. Box 10414 Jerusalem, ISRAEL
offerd@cs.huji.ac.il

Abstract. Most information systems, which perform computerized searches of textual databases, deal with the need to display a list of documents which fulfill the search criteria. The user must choose from a list of documents, those documents which are relevant to his search query. Selection of the relevant document is problematical, especially during searches of large databases which have a large number of documents fulfilling the search criteria. This article defines a new hierarchical tree which is made up of three levels of display of search results. In a series of previous studies (not yet published) which were carried out at the Hebrew University in Jerusalem, the influence of information (within the documents) displayed to the user was examined in the framework of a list of responses to questions regarding user satisfaction with the method and the quality of his choices. In the present study, in addition to the information displayed in the list, information on the contents (subject) of the document was also displayed. The study examined the influence of this additional information on search time, user satisfaction and ease of using the systems.

1 Introduction

1.1 Background

In information systems based on existing textual databases, there are two main methods for displaying lists of documents which fulfill search criteria. In the first method, which we shall refer to as method A, only the titles of the documents are displayed, where the user has the possibility of going to the document itself and deciding as to its relevance. In the second method (B), the titles of documents are displayed together with a number of lines from the beginning of the document. This is done in order to provide a certain indication to the user regarding the relevancy of the document. The assumption of this method (which constitutes, in practice, an extension of the first method) is that the beginning of the document is characteristic of its contents. In scientific documents, the beginning of a document contains an abstract or introduction characterizing the subject of the document. At the beginning of the document, details identifying the writer and his place of work are presented and can be used to disclose something regarding the contents of the document and its quality.

Besides these two methods, a third method for the display of lists of documents from databases was designed and investigated by this writer. This method (C), includes the display of a number of lines from the relevant paragraph, which fulfills

the search condition [5][18]. These lines appear under the titles of the documents. The point of departure of this method is that most of the databases in use today are not keyed in advance and documents which are not scientific do not undergo a well-organized control process. In the light of these assumptions, the information that has to be delivered to the user must be the most relevant that is possible to find in the document, based on the search criteria of the search query, and not based on a fixed location in the document. For these reasons, the most relevant location is found in the body of the document, located in the lines which fulfill the search condition and which were the reason that the document was included in the list of responses. The results of the investigation which examined this method showed, in a clear manner, that users of the system feel that this method provides greater ease of use than do other methods, are satisfied with it and expect that this is how information systems should behave. No significant difference was found in search time between this method and the other accepted methods [6].

1.2 Previous Research

In the past, a number of studies were conducted regarding the display of search results for textual databases based on a list of documents. The following are the principal studies carried out.

In a study of types of interfaces conducted by Hertzum and Frokjaer, the advantage to the user was examined for various interfaces. The first interface displayed to the user a line of linkages which constituted the table of contents of the database. User navigation within the database was carried out through a dialogue between the various linkages, where every linkage represented an additional level in the hierarchy. The second interface enabled the user to carry out a Boolean search on a database, where the search results were displayed as a Venn diagram. In this interface, the number of the documents fulfilling the conditions were displayed in the diagram and the user could see, at a single glance, all the documents fulfilling all of the search criteria. The third interface was a combination of the two previous interfaces. In contrast to the interfaces described above, the database was printed and a separate group performed the same tasks, working from the printed material and without using a computer. The results of the study showed that there was no significant difference between the quality of the results obtained by the different methods. The navigation approach was found to be faster than the Venn approach, but both were slower than working from a printed guide! [12]

In another study on the subject of display of information from a given database, three interfaces were compared. In the first interface, the complete index of the database was displayed. This included three hierarchical levels of titles without any condensation, whatsoever. In the second interface, only the upper level titles were displayed, while the user could mark the desired level and receive a display of the level subsidiary to the one requested. In the third interface, the screen was divided into three horizontal windows. In the upper window, only the upper level titles were displayed; in the middle window, middle level titles were displayed; and in the bottom window, third level titles were displayed. The experiment included the performance of a variety of tasks using the different interfaces. The results of the study showed that the highest speed was attained with the second and third methods.

For tasks requiring a great deal of navigation, the third method was found to have an advantage over the second one [4].

Landauer, in his book "The Trouble with Computers", presents an additional approach, called "Fisheye". In this approach, a list of documents from a given database is displayed, where the list of documents constitutes a kind of table of contents of the database. Broad and in depth detail is given to the area of the document which interests the user and on which he is focused; and as one moves away from the focus of interest, the number of levels displayed on the screen is reduced, so that, in effect, we move up in the hierarchical level of the database [14].

Another experiment involved the development of an electronic book designed from a user interface for displaying text. The project, called Superbook, was used in a number of experiments in which a comparison was made between information retrieval from a printed book and from its electronic version. The user interface in Superbook contained three vertical windows on the screen. The left hand window displayed the table of contents of the document using the "fisheye" method. In the center window the user could type the required words, and in the right hand window, the relevant texts were displayed. In a study carried out by Egan and his colleagues, it was found that for most user tasks, Superbook produced better results in finding the required textual segments, and in the use of texts which were not edited in advance and, also, in the time devoted to finding the answer [8].

Another experiment in displaying textual information was made by Pirolli and his colleagues. A document display system, called Scatter/Gather Browsing, displays information in 10 corresponding windows in two columns. The windows themselves constitute a "cluster" which gathers a number of documents having words with associated significance. Under these associated words, the titles of the first three documents are displayed. The user may widen the information (scatter) or reduce the information (gather) by means of appropriate commands. The experiment compared the display of information using this method with that of a different system (SimSearch), which made it possible for the user to introduce a series of desired words, following which the system produces a small set of documents which fulfill the search criteria. The results of the study showed an advantage to the Scatter/Gather method for tasks which look for a specific document [21]. Another advantage of using this approach was found when there was a need to leaf through large, unknown databases. Expansion through producing groups of documents which have a common subject was shown by Pitkow and Pirolli. According to their method, if two different documents (a and b) are cited by a third (c) it may be supposed that those two documents (a and b) are related. The more frequently the group of documents is cited, the greater the possibility of defining the level of interconnection between them and, thus, of gathering them together during display [22]. An additional proposal for designing a results window is given by Shneiderman in his book "Designing the User Interface". According to his proposal, control over a number of components should be left in the hands of the user of the system, so that he can define them in the way that he finds easiest. As an example of these components, he brings the number of documents that will be received as a result of the search criteria, those data fields which will be displayed in the results window, the document sort in the list (a-b, according to dates, etc) and the method for gathering the results into clusters (according to the value of certain characteristics, such as subject) [23].

It should be noted that there are a number of other studies for displaying the results of textual database searches that deal with methods that are not based on a list of

documents, but on graphic elements which are intended to demonstrate the relevance of the documents received and the relationship between them. This study will not deal with approaches, but will only note their existence for the interested reader: Projects WebBook and The Web Forager [1], Clangraph [26], Butterfly [19], TileBars [10], Value Bars [3], Cat-a-Cone [11], InfoCrystal [24], VIBE System [20], Document Network Display [9], AspInquery Plus System [25], SPIRE [27], Kohonen Self-Organizing Map [17], WEBSOM [13], Envision [19a] and ET-MAP [2].

1.3 Problems in Finding Relevant Documents

The performance of a large number of searches using a number of text retrieval systems showed that the larger the size of database, the greater the chances of error or “noise”. There are several reasons for this:

1. The size of the database directly influences the amount of answers which are received.
2. The size of the database directly influences the spread of answers (on the assumption that the database is not dedicated to a certain subject).
3. Large databases (especially opened ones) tend not to give accurate results to their searches. It is a known phenomenon that the performance of the same search strategy may yield different results or that the use of the identical Boolean strategy, from the standpoint of search contents, will produce different results.
4. In large databases, there is a tendency, because of commercial competition, to display the largest number of possible results, and doing so at the expense of accuracy. For example, a search for a certain word will produce, in one database, the documents which include that particular word exactly as it is written in the search statement, as opposed to a different database which will automatically “broaden” the word and will include documents containing the word itself, as well as its derivatives. The central problem is that the majority of systems do not publish their defaults, and even if the user succeeds in learning them, he is not given the possibility of controlling them.

A different problem which also contributes to “noise” is related to the fact that some of the documents which fulfill the search criteria deal with completely different areas than the subject area that interests the user. This is possible in several situations.

1. In every language there is the phenomenon of “homograph”, where words have identical spellings, but different meanings. In this situation, when using a standard search for a particular word, the answers received will relate to documents in which the word is found. If the word is a homograph, the documents received will come from variety fields in which this word exists.
2. Also, words having a single meaning have a natural diversity of uses. A word may be a central word in a document and constitute the subject, or at least part of the subject, of the document. The same word can have the same meaning, but with a completely marginal connection to the contents of the document.
3. Words can be part of a combination of words which change their meanings. The word “air” has a certain connection to the field of chemistry, while the word combination “air observation” is related to aviation, even though the word “air” has the same meaning in both cases.

The examples presented above reveal an almost vital need to deal with the subject/content of the document or the subject/content of the database from which the document is drawn, as well as the need to refine searches in large databases. It should be possible to respond to these needs by dealing with those parts of the data which are to be displayed to the user at the time he receives the results of the search he requested.

2 Techniques for Displaying Search Results

Together with various studies on what information from the document should be displayed in the list intended for user examination, the possibility of displaying information based on the contents of the document or the database was examined. This was done in order to provide the user with immediate and concentrated information for making a decision as to whether or not a given document is relevant to the search query. The point of departure of these studies, like that of the research described above [6], is the desire to give the user of the system the ability to decide what constitutes a relevant document to his search query, without having to read the entire document. In order to enable him to perform this task, vital information must be given to the user and it must be displayed in a way that permits him to perform the task.

In this section, we will define a three level hierarchical tree for the display of the search results (see Figure 1). This definition is new and is one of the contributions of this article.

The upper level includes two categories for displaying the search results. The first category is based on textual techniques, while the second category is based on graphic techniques which include the use of graphic objects to describe the documents found in the results and the relationship between them. There are other interesting combinations of graphic and textual techniques in use and, in fact, most graphic user interfaces actually visualize textual elements. These, however, will not be discussed, here. This study deals with the textual category and, therefore will not expand upon graphic methods. In the textual category, we distinguish two methods. One is a list of the results based on information within the document and the other is a list of results based on information which is external to the document.

2.1 List of Results Based on Information within the Document

In this category, use is made of a number of methods which enable designing a list of results where each item in the list is based on information from one of the methods or a combination of them. The following is a description of the various methods.

2.1.1 Significant Sentences

It is possible to find significant sentences within the document which serve to describe it. It is possible to select descriptive sentences based on defined paragraphs in the documents, such as: abstract, introduction, and summary. In addition, it is possible to use sentences relevant to the search query. These sentences include the

search words which were requested by the user and which served as the basis for selecting the particular document that fulfilled the search criteria [18], [5].

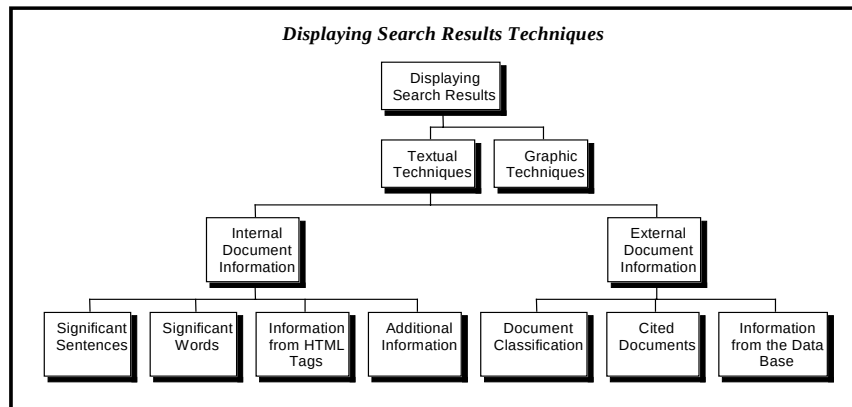


Fig. 1. Display techniques for search results (UTECDSR)

2.1.2 Significant Words

It is possible to use significant words within the document which describe it, such as key words or frequently used words. It is possible to make use of key words which have been prepared by the author of the document or by automatically producing them at the onset. Using frequently used words (by scrolling through the Stop List) can produce similar results, but this is not exact in all cases.

2.1.3 Information from HTML Tag

In HTML documents on the Internet, it is possible to make use of the language tags to find information about the document. One can locate, for example, the names of paragraphs or sub-headings, by using the <H> (Header) tags. It is possible to use <META> tags, which contain information about the document as it was recorded by the writer of the document. These tags can facilitate locating information that constitutes an abstract of the document, key words, etc. There is, of course, a certain amount of “noise” in these tags, because of the use commercial bodies make of these tags to raise the rating of the document in their various search engines.

2.1.4 Additional Information

Additional information found within the document itself can be used, for example, in a document which contains citations of other documents. It is possible to use the titles of the documents cited on the assumption that they deal with a closely connected subject area and can suggest something about the contents of the document, itself.

2.2 List of Results Based on Information External to the Document

In this category, use is made of a number of methods which enable displaying a list of results in which the information displayed for each item on the list is based on the subject area of the document. It is possible to use any one of the methods separately or in combination.

2.2.1 Classification of the Document

In this category, use is made of a number of methods making it possible to display the category to which the document belongs. It is possible to use search engines which manually define the category to which the document belongs (for example YAHOO). It is possible to produce a category with the help of computerized algorithms and it is possible to determine the subject area to which the document belongs through clustering operations for all of the search results [28] [16][15].

2.2.2 Cited Documents

In this category, it is possible to determine subject area of the document based on external documents in which it is cited. The assumption is that a document is cited by the another document when they have common subjects. Finding the cited documents can be done directly, based on the Internet, or it can be done using a dedicated database, such as the Science Citation Index. When searching for cited documents, it is possible to used their titles or, alternatively, the names of the cited paragraphs.

2.2.3 Information from the Document's Database

Information from the database in which the document is found can give an indication of the subject area of the document in various ways. When the database is dedicated, it is possible to obtain information about the subject of the database from within it. One can find additional documents in the database and use their titles to determine the subject area of the database. In addition, it is possible to critically examine the name of the directory in which the documents are stored and, perhaps, learn from it something about the database itself.

3 Format of the Study

In order to test part of the model presented above, a study was conducted in which a selection of the components of the tree were examined. The components dealt with included the display of significant sentences, while performing a comparison of the sentences which were described as relevant and the use of significant words in comparison to the relevant sentences [7][6].

3.1 Question of the Study

The question examined in the study was: Does the addition of information reflecting the contents of the document, such as keywords, improve the search process in the opinion of the user and does it shorten the search time?

3.2 Hypotheses of the Study

1. The addition of key words to the information displayed in the list of search results will shorten the search time, when compared to information displayed without the key words.
2. The addition of key words to the information displayed will improve the feeling of ease of use for the users of the system, when compared to the same information without key words.

3. The addition of key words to the information displayed will improve the user's sense of satisfaction, when compared to the same information without key words.

3.3 Format of the Experiment

The layout of the experiment included having a group of users perform various search tasks using different methods. Each one of the groups made use of each one of the methods (5 tasks all together). The layout of the experiment included building three information databases where each database included 30 documents in the English language for a given subject. All of the documents in the database were collected from the Internet and their contents were relevant to the search task. A portion of the documents were only textual and a portion of them included pictures as they appeared in the original. Use of the information database was intended to avoid "noise" during the experiment and permit all of the participants in the experiment to receive an identical number of documents for each search task, as well as the same documents. The number of answers for each task was identical (30). The databases which were built were general and intended for simple search tasks, such as the birthday of the physicist Newton, finding the telephone number of a flower shop in a certain city which makes deliveries to different city, the amount of money that the Americans paid to the Russians for the purchase of Alaska, etc. The tasks were given to the participants in writing and they were required to find the answers from within the database. In addition, a software program which performed a simulation of a search engine was developed. It worked opposite the various databases during performance of the search. This was done with one interface for all the trials and the output of results was different among the trials themselves. The various search tasks were defined for the users participating in the experiment, where each task had one factual answer and one document in the database which contained the correct answer. The assignment of the tasks to different interfaces was done randomly, so that each user performed different search tasks on different interfaces and, thus, in practice, experienced using all the interfaces. The methods were presented counterbalanced across subjects. All the users used the Netscape (ver. 4) browser, so that the interface for the experiment would be uniform. Gathering the results was done through a computerized activities log and by participants filling out feedback sheets. The data which was received from the experiments was derived from two categories: objective data and subjective data. The objective data included search times, as well as the correctness of the response to the task. The subjective data included feelings and preferences as they were recorded on the response sheet.

3.4 Study Population

All of the participants were students studying for a Master's Degree in the School Of Business Administration and in the School of Library and Information Science from the Hebrew University in Jerusalem. There were 128 participants in the first experiment and 51 in the second. In both experiments, the participants performed five search tasks. The participants in both experiments were alike (age, subject of learning, using computer experience etc.). The time period between the two experiments was 9 months. The selection of the population was random. The selection of the task for the method was, also, random.

3.5 Analysis of Results

The analysis of the results was made using a variety of statistical tools. Anova tests examined the degree of compatibility or disparity of the different methods. Since each participant used all of the methods (three, altogether), it was possible, by means and Duncan Grouping, to select which methods was preferred in the opinion of the users. Another statistical tool for examining the ranking of pairs of methods was the Mcneamar symmetrical matrixes test. In addition, the P test for determining rankings of results was used, as was, of course, regular statistical analyses of averages, standard deviations, etc.

3.6 Methods of Displaying Information Which Were Studied

In the first experiment, the following possibilities were displayed:

1. A list on which only the document title of each document was displayed (Method A).
2. A list on which the document title of each document which was displayed + the three first lines from the beginning of the document (Method B).
3. A list on which the document title of each document was displayed + three relevant lines which fulfill the search criteria taken from the body of the document (Method C).

In the second experiment, the same possibilities were displayed, except that additional information was added concerning the subject of the document which, in this experiment, included its list of key words (see figure 2).

In this experiment, the subject/contents of the document was defined as a series of key words from the document. To facilitate the experiment, use was made of a function in the Word word processor (automatic abstract, characteristics, key words) which produces the key words of a document.

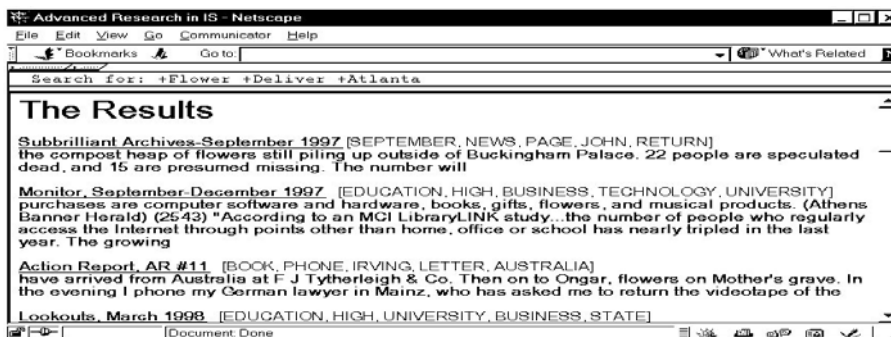


Fig. 2. User Interface of the system in method C

4 Results

We found (using Anova test) that there is significance difference between the methods on both experiments ($p < 0.0001$).

Hypothesis 1 - Search time

In Figure 3, the search times for the different methods (A, B, C) according to the different search tasks (task 1, task 2 task 3 from the 5 tasks) are shown. As the table shows, search time was significantly reduced in most cases. Using method C, the time was shortened by as much as 57% (average 33%), with Method B, by as much as 51% (average 17%) and with method A, by as much as 31% (average 14%). We found that in a simple tasks method A was faster. We distinguish between simple task to heavy task by checking the average time to complete the task and by measuring the numbers of error results between the tasks. We think that in a simple tasks the solution was found in the search result list (in the titles) and in method C the user spent time to read the abstract.

Hypothesis 2 - Ease of Use

The feedback question which examined this hypothesis was “To what degree is there a feeling of ease with each of the different methods?” The values assigned to the answers were from ‘1’ to ‘10’, where ‘1’ represents “not at all easy” and ‘10’ represents “very easy”. The results obtained show that there was an increase in the percentage of ease, as defined by the users, for all of the methods (see figure 4). With Method A, there was a 66% increase in the feeling of ease (7.6 with key words, as opposed to 4.6 without key words, Std dev=1.66, $P=0.0001$). With Method C, there was a 12% increase (9.0 with key words, as opposed to 8.0 without key words, Std dev=1.88, $P=0.0044$). With Method B, there was a 29% increase in the feeling of ease (7.9 with key words, as opposed to 6.1 without key words, Std dev=1.40, $P=0.0001$).

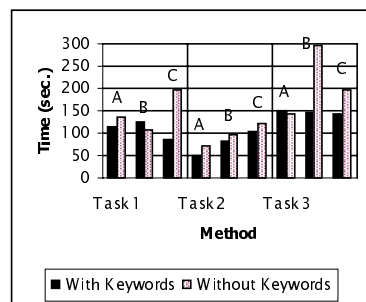


Fig. 3. The relationship between search times when using key words and not using them with different methods for different tasks

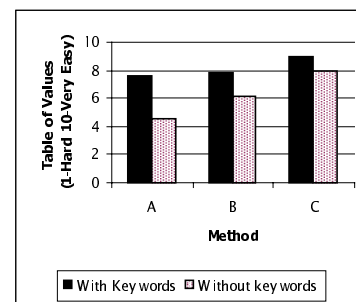


Fig. 4. User analysis of ease using different methods (and with respect to the use of key words and without them)

Hypothesis 3 - Satisfaction

The feedback question which examined this hypothesis was: “To what degree were you satisfied with each of the different methods, supposing that the search engines were used with the proposed system?” The values of the answers were from ‘1’ to ‘10’, where ‘1’ represents “not at all” and ‘10’ represents “very satisfied”.

The result obtained showed that there was an increase in the percentage of satisfaction, as defined by the users, with all the methods (see figure 5). With Method A, there was a 59% increase in the feeling of satisfaction (6.5 with key words, as opposed to 4.1 without, Std dev=1.96, $P=0.0001$). With Method C, there was a 12%

increase (8.9 with key words, as opposed to 7.9 without key words, Std dev=1.52, $P=0.0022$). With Method B, there was an 8% increase in the feeling of satisfaction (7.4 with key words, as opposed to 6.8 without key words, Std dev =1.52, $P=0.0251$).

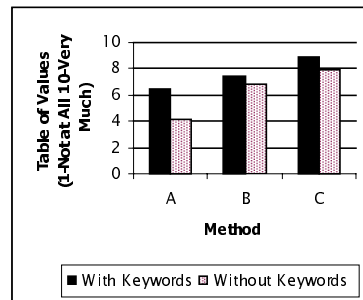


Fig. 5. User satisfaction using different methods (and with respect to using key words and without them)

5 Conclusions

5.1 Direct Conclusions from the Assumptions of the Study

1. The addition of key words to the information displayed on the list of the search results will shorten the search time, in contrast to information displayed without key words.
2. The addition of key words to the information displayed in the list of search results will improve the user's feeling of ease, in contrast to the same information without key words.
3. The addition of key words to the information displayed in the list of search results will improve the user's feeling of satisfaction, in contrast to the same information without key words.

5.2 General Conclusions

Based on an analysis of the results from a statistical standpoint, it is possible to see that:

1. There is a clear advantage in using key words in everything related to shortening search time. The advantage is clearest with Method C.
2. Method C is the fastest for complex search tasks. Method A was faster for simple tasks.
3. The most pronounced improvement in using key words, in terms of everything relating to a feeling of ease and a feeling of satisfaction, is found with Method A. Method A is the method which displays the smallest amount of information (only the title of the document). As such, all additional information is very significant when compared to the other methods.
4. Method A is the least recommended for displaying search results (feeling of ease, satisfaction and search time in complex tasks). On the other hand, if there are constraints which necessitate using of this method, the addition of key words to this

method will significantly contribute to shortening search time and improving the ease of use the system.

5. Method C is considered the easiest to use with or without key words (see figure 4). The additional information which was found to be effective was key words and relevant sentences fulfilling the conditions of the search. One must indicate that the use of relevant sentences is always possible, as they are the cause of the document being entered on the list of results. Technically, this task is not complicated for “full text” databases, in which all of the words are included in the database, and locating the sentence which includes the search words is possible with almost all search engines. The use of key words will be easier for documents which include key words in the original or, alternatively, which produce key words based on existing algorithms. The main drawback is the pressure this puts on the system.

6 Summary

This article proposes a new definition for a three level hierarchical tree for displaying search results. Together with the definition, a study was conducted which examined the degree to which the addition of information to the list of search results of a textual database constitutes an advantage during the search process. The additional information which was examined in this study related to key words from the document which were displayed in the list of search results and lines in search context that added to document's titles. For all the criteria examined, the list which included key words received higher marks than those lists which were displayed without key words. The hypotheses of the study which were supported, showed an advantage in the following areas: shortening of search time, feeling of ease with the method proposed and a feeling of satisfaction. It should be noted that a greater relative advantage was systematically recorded when displaying key words with Method C, as compared to doing so with Method B. This tells us that combining search method C with key words is, in the user's opinion, optimal. We believe that using this method with the search engines which currently exist on the Internet can significantly improve the ability of the user to easily and comfortably find information on the Internet.

7 Suggestions for Additional Studies

An additional study could focus on a number of subjects which this study related to, but did not examine:

1. The examination of the advantage of additional characteristics related to the document, which are not key words: for example, the categories to which the document is related, use of frequently used words, use of HTML tags, etc.
2. The examination of the advantage of additional characteristics relating to the document's environment: for example, performance of access to existing links in the document and displaying their titles.
3. Performing a parallel study designating a specific professional population (in contrast to the present study, which was general in nature), which will include

professional tasks for the same population and examine the validity of the assumptions for different professional realms.

4. An examination of what is the most valuable factor in the opinion of the user: information from within the document or information from the document's environment.

Acknowledgements

I would like to thank Eliezer Lozinskii for his useful remarks and Avner Ben-Hanoch for his assistance in gathering the data in the research. My thanks also to Tomer Drori for his help in the graphic display of data.

References

1. Card, K., Robertson, G., York, W. The WebBook and the Web Forager: An information workspace for the World-Wide Web. *Proceedings of CHI '96*, ACM, (1996), 111-117.
2. Chen, H., Houston, L., Sewell, R., Schatz, R. Internet browsing and searching: User evaluations of category map and concept space techniques. *Journal of the American Society for Information Science*, 49(7), (1998), 582-603.
3. Chimera, R. Value bars: An information visualization and navigation tool for multi-attribute listings. *Proceedings of CHI'92*, ACM, (1992), 293-294.
4. Chimera, R., Shneiderman, B. An exploratory evaluation of three interfaces for browsing large Hierarchical tables of contents. *ACM Transaction in Information Systems*, ACM, 12, 4, (1994), 383-406.
5. Drori, O. The user interface in text retrieval systems. *SIGCHI Bulletin*, ACM, 30, 3, (1998), 26-29.
6. Drori, O. Displaying Search Results in Textual Databases, *SIGCHI Bulletin*, ACM, 32, 1, (2000), 73-78.
7. Drori, O. Using Text Elements by Context to Display Search Results in Information Retrieval Systems, Information Doors (a workshop proceedings held in conjunction with the ACM Hypertext 2000 and ACM Digital Libraries 2000 conferences), San Antonio, Texas, USA., (2000), 17-22.
8. Egan, E. et al. Behavioral evaluation and analysis of a hypertext browser. *Proceedings of CHI '89*, ACM, (1989), 205-210.
9. Fowler, H., Fowler, A. and Wilson, A. Integrating query, thesaurus, and documents through a common visual representation. *Proceedings of SIGIR '91*, ACM, (1991), 142-151.
10. Hearst, M. TileBars: Visualization of term distribution information in full text information access. *Proceedings of the CHI '95*, ACM, (1995), 59-66.
11. Hearst, M., and Karadi, C. Cat-a-Cone: An interface for specifying searches and reviewing retrieval results using a large category hierarchy. In *Proceedings of SIGIR '97*, ACM, (1997).
12. Hertzum, M., Frokjaer, E. Browsing and querying in online documentation: A study of user interface and the interaction process. *ACM Transactions on Computer-Human Interaction*, ACM, 3, 2, (1996), 136-161.
13. Kohonen, T., Exploration of very large databases by self-organizing maps. *Proceedings of the IEEE International Conference on Neural Networks*, vol. 1, (1997), 1-6.

14. Landauer, K. *The trouble with computers - Usefulness, usability and productivity*. Cambridge MA: MIT Press, (1995).
15. Larkey, L. Automatic essay grading using text categorization techniques. *Proceedings of SIGIR '98, ACM, (1998)*, 90-95.
16. Larkey, L., Brucecroft, W., Combining classifiers in text categorization. *Proceedings of SIGIR '96, ACM, (1996)*, 289.
17. Lin, X. A self-organizing semantic map for Information retrieval. *Proceedings of SIGIR '91, ACM, (1991)*, 262-269.
18. Luhn, H., Keyword in Context Index for Technical Literature, *American Documentation*, XI (4), (1960), 288-295.
19. Mackinlay, D., Rao, R., Card, K. An organic user Interface for searching citation links. *Proceedings of CHI '95, ACM, (1995)*.
- 19a. Nowell, T. et al., Visualizing Search Results: Some Alternatives To Query-Document Similarity, *Proceedings of SIGIR '96, ACM, (1996)*, 67-75.
20. Olsen, A., Korfhage, R., Spring, B., Sochats, M. and Williams, G., Visualization of a document collection: the VIBE system. *Information Processing and Management*, 29(1): (1993), 69-81.
21. Pirolli, P. et al., Scatter/gather browsing communicates the topic structure of a very large text collection. *Proceedings of CHI '96, ACM, (1996)*, 213-220.
22. Pitkow, J., Pirolli, P. Life, death, and lawfulness the electronic frontier: *Proceedings of CHI '97, ACM, (1997)*, 213-220.
23. Shneiderman, B. *Designing the user interface: Strategies for effective human-computer interaction*, 3rd ed. Reading, Massachusetts: Addison-Wesley, (1998)
24. Spoerri, A. InfoCrystal: A visual tool for information retrieval and management. *Proceedings of Information Knowledge and Management (CIKM'93)*, (1993), 150-7.
25. Swan, R. and Allan, J., Aspect windows, 3-D visualizations, and indirect comparisons of information retrieval systems. *Proceedings of SIGIR '98 , ACM, (1998)*.
26. Terveen, L, Hill, W., Finding and visualizing Inter-site Clan graphs. *Proceedings of CHI '98, ACM, (1998)*, 448-455.
27. Wise, J.A., Thomas, J.J., Pennock, K., Lantrip, D., Pottier, M., Schur, A. and Crow, V. Visualizing the non-visual: spatial analysis and interaction with information from text documents. *Proceedings of the IEEE Information Visualization symposium*, (1995), 51-8.
28. Zamir, O., Etzioni, O. Grouper: A dynamic clustering interface to Web search results WWW8 *Proceedings, Toronto: WWW, (1999)*.

Interactive-Time Similarity Search for Large Image Collections Using Parallel VA-Files

Roger Weber, Klemens Böhm, and Hans-J. Schek

Institute of Information Systems
ETH Zentrum, 8092 Zurich, Switzerland
{weber,boehm,schek}@inf.ethz.ch

Abstract. In digital libraries, nearest-neighbor search (NN-search) plays a key role for content-based retrieval over multimedia objects. However, performance of existing NN-search techniques is not satisfactory with large collections and with high-dimensional representations of the objects. To obtain response times that are interactive, we pursue the following approach: it uses a linear algorithm that works with approximations of the vectors and parallelizes it. In more detail, we parallelize NN-search based on the *VA-File* in a *Network of Workstations* (NOW). This approach reduces search time to a reasonable level for large collections. The best speedup we have observed is by almost 30 for a NOW with only three components with 900 MB of feature data. But this requires a number of design decisions, in particular when taking load dynamism and heterogeneity of components into account. Our contribution is to address these design issues.

1 Introduction

Images are an integral part of digital libraries [14, 10, 6]. Given large image collections, a pressing problem is to find images that cover a given information need. Due to the huge amount of data, browsing-based approaches and traditional retrieval methods are not feasible. E.g., keyword-based techniques are too expensive, requiring too much manual intervention. In contrast, a *content-based information retrieval system (CBIR system)* identifies the images most similar to a given query image or query sketch, i.e., it carries out a *similarity search* [7].

A common approach to similarity search is to extract so-called *features* from the objects, e.g., color information. A feature typically is represented by a point in a high-dimensional data space. Then CBIR means finding the *nearest-neighbor (NN)* to a given query point in that space. However, performance of many approaches is a problem if the feature data is high-dimensional: search time is linear in the number of data objects (cf. Related Work). Thus, as soon as NN-search is I/O-based, it is expensive. Absolute response times are not satisfactory when collections are large. We are not aware of approaches that can safely be deployed in a digital library (DL) that is operational.

Consequently, linear approaches that explicitly inspect all feature vectors, sometimes shrugged off as 'naive approaches' or 'brute force', are competitive in

many situations. The *VA-File* [16] is a linear approach that uses approximations of the vectors instead of the full vectors for most computations. This reduces disk I/O and improves performance by a factor of 4 as a rule of thumb. On the other hand, it obviously does not solve the problem that search times grow linearly with the number of objects.

Having said this, it is obvious to evaluate the potential of parallel NN-search, i.e., using a *Network of Workstations (NOW)* to carry out the search in parallel. With parallel tree-based NN-search [9, 1, 12], the speedup is only sub-linear (e.g. a factor of 6 with 16 disks). On the other hand, parallelizing the linear approach is straightforward, and speedup is expected to be roughly linear in the number m of components: given a query, each component iterates through a partition of the feature data, and the so-called *coordinator machine* then computes the overall query result. While this looks simple in principle, difficulties and design questions arise when it comes to implementing a DL component that is operational. This imposes further requirements, e.g., coping with load dynamism and heterogeneity of the components or tolerating failures.

The main contribution of this article is 'on the engineering level', i.e., we come up with a comprehensive list of requirements or design objectives regarding parallel VA-Files, and we identify design decisions and relate them to the objectives. Having conducted extensive experiments to evaluate the alternatives, we report on the most interesting results. With regard to experiments, the most impressive number we have observed was a speedup of 30 when moving from one to three components. From another perspective, searching a gigabyte of feature data lasts only around one second. Thus, our results allow to meet the difficult requirement of interactive-time similarity search.

2 Related Work

NN-search in high-dimensional vector spaces. Tree-based methods for NN-search partition the data space according to the distribution of the data. The idea is to prune the search space so that the access cost is only logarithmic in the number of objects. Some of these methods use the absolute position of objects for clustering [8, 2], others use distances between objects [4]. While these methods provide efficient access to low-dimensional data, performance degrades rapidly as dimensionality d increases [2, 1, 16]. As a consequence, it looks much more promising to deploy scan-based algorithms for NN-queries.

The VA-File. The idea behind the VA-File is to reduce the amount of data read during search [16, 15]. The VA-File consists of two separate files: the vector file containing the feature data, and the approximation file containing a quantization of each feature vector. The nearest neighbor of a query is found similar to signature techniques [11, 5]: a first phase (filtering step) scans only the approximations. While doing so, it computes bounds on the distance between each data point and the query. These bounds suffice to eliminate the vast majority of the vectors, and a very small set of candidates remains, both for synthetic and

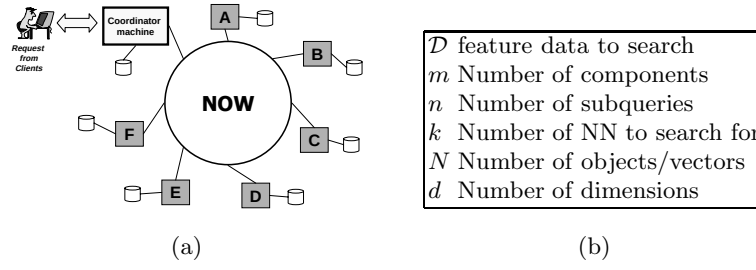


Fig. 1. (a) Parallel NN-search in a NOW. (b) Notational summary.

natural data sets. A second phase (refinement step) accesses some of the vectors corresponding to these candidates to determine the nearest neighbor. Next to the nearest neighbor, the second phase only visits very few additional vectors, typically between 5 and 10. Since the VA-File has an array-like structure, each data item has an absolute position in both the vector and the approximation file.

Parallelization of tree based methods. [9] presents a parallel version of the R-tree. An important issue is to assign the children of a node to different disks. A promising strategy is to measure the probability that a query retrieves two child nodes. It works well in the two-dimensional case, but the improvement is not linear. Similarly, [1] performs well as long as dimensionality is small. Given a multi-disk architecture and a parallel R-tree, [12] investigates how many nodes should be fetched from the disks in parallel during each step of query evaluation. They describe only experiments up to $d = 10$ and do not give absolute numbers.

3 Architecture for Parallel VA-Files

In the following, we first give an overview of the architecture and list design objectives. We then discuss the design alternatives and the respective tradeoffs in detail. Table 1 (b) summarizes the notation.

3.1 Architecture and Coordination

Figure 1 (a) depicts the architecture for parallel VA-File based NN-search in a NOW: a *coordinator* receives a similarity query from a client, divides it into n subqueries and distributes them among the m components of the cluster. It then assembles the results of the subqueries and returns the result. Subsequently, we discuss the design objectives and summarize the design decisions.

Design Objectives. When parallelizing the VA-File, a number of (conflicting) requirements arise:

- *Dynamism of components.* There are two facets to this requirement: if the components do not only do NN-search, but run other applications as well, workloads may change significantly in short time. Similarly, the system should take advantage of new components without manual intervention. *Failure tolerance* must be a feature as well.
- *Heterogeneity of components.* This requirement is important in practice: in the vast majority of cases, one will not encounter the situation that all workstations are identical. Coping with heterogeneity should lead to a much better exploitation of the resources available.
- *Dynamism of data.* Typically, the image collection changes over time, i.e., updates take place. In a nutshell, the requirement is that updating the Parallel VA-File does not become too costly in the presence of replication.
- *Ensuring good query performance.*
- *Economic usage of main memory and disk space.* This requirement might be of secondary importance, given the recent drastic price drops.

Design Decisions. We have identified the main design alternatives and the respective tradeoffs. The alternatives also serve as dimensions of the space of our experiments to follow. We summarize them as follows: (1) data placement, (2) partitioning of subqueries, and (3) exploitation of memory resources. However, these dimensions are not fully orthogonal, as we will explain.

3.2 Data Placement

With regard to data placement, we need to distinguish between the approximation data and the full vector data. It is subject to investigation (a) if the approximation data should be replicated, or each component should hold only a portion of the approximation data, and (b) if each component should hold a full copy of the vector data, or only one component maintains such a full copy.

To be able to comment on the alternatives with regard to (a), we discuss what the coordinator does with a given query: it generates n subqueries for the query. Each subquery inspects another partition of the data set. Each partition is an interval over the list of data objects. The coordinator must determine the bounds of these intervals. We use the following notion: Let $NN(\mathcal{D}, k)$ denote the NN-query over $\mathcal{D}$, and let $NN(i, \mathcal{D}, k)$ denote the i -th subquery. The coordinator must then determine bounds b_i for $0 \leq i \leq n$ such that $b_0 = 0$, $b_n = N$, and $\forall i : b_i < b_{i+1}$. Then interval $[b_i, b_{i+1}[$ identifies the i -th partition. The computation of these bounds depends on the placement alternatives for the approximation data: with replication, the coordinator can choose any bounds that meet the above conditions. In the other case, this is obviously not feasible.

Each component computes candidates for its subquery. Now consider Item (b). Depending on the placement of the vector data, it identifies the k best data objects by accessing its local disk or by accessing a central component.

Obviously, the replication alternatives better cope with the requirement 'dynamism of components', but they incur higher update costs. Replicating the full

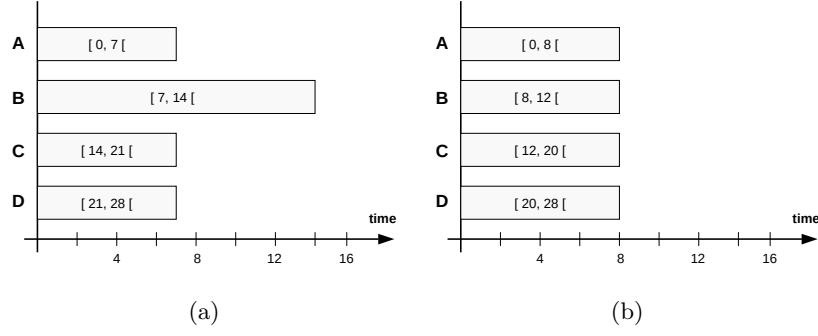


Fig. 2. Query evaluation in a heterogeneous environment with (a) equal partition sizes and (b) individual partition sizes (**B** is half as fast as the other components).

vector data among the components is potentially better with regard to query performance, as it incurs lower communication costs. On the other hand, this is not exactly economic usage of resources.

3.3 Partitioning of Subqueries

So far, we have left open how to choose the number of subqueries and their bounds in case of replicated approximation data. Without replication, there is no liberty in choosing the subqueries. A first decision is whether the partitions have the same size or not (*equi* vs. *individual*). With individual, the size of each subquery reflects the capability of the corresponding component and each component evaluates one subquery. With *equi*, the picture is more differentiated: we can have many small partitions and more subqueries than components. The idea is that each component works off a different number of subqueries, again depending on its capabilities.

In a homogeneous environment without component failures, the computation of b_i is straightforward and identical for *equi* and *individual*: the partitions of the subqueries are of equal size, and the number of subqueries equals the number of components, i.e. $n = m: \forall 0 \leq i \leq n : b_i^{eq} = i \cdot N/n$. But in general, the nodes of a NOW differ in terms of CPU power, memory and disk capacities. Thus, the execution times of subqueries of identical partition sizes may differ as well. Figure 2 depicts the evaluation of a query in a heterogeneous environment. Let $N = 28$. The interval within the bars denotes the partition of the subquery corresponding to the component. Assume that components A, C and D are twice as fast as component B. With equal partitions, the search cost is 14 since B needs twice the time of the other components to evaluate its subquery (cf. Figure 2 (a)). Note that during the last 7 time units only B is doing work related to the query. On the other hand, if the size of the partition of the subquery of B is half the one of the other components, as depicted in Figure 2 (b), the duration of the search is only 8. This is the optimum.

More formally, we can compute the partition bounds with individual as follows: the coordinator knows the time to search a partition of length L with each component. Let s_j denote the number of data items component j can search within one time unit. A large value of s_j indicates that j is fast (or that its workload is low). To reach optimality, the partition size l_j of the subquery of component j must be proportional to s_j :

$$l_j = N \cdot \frac{s_j}{\sum_{j=0}^{m-1} s_j} \quad (1)$$

If the number of subqueries equals the number of components ($n = m$), the bounds of the partitions are:

$$b_0^{dyn} = 0, \quad \forall 0 < i \leq n : b_i^{dyn} = \sum_{j=0}^{i-1} l_j \quad (2)$$

We can deploy this computation scheme if the workloads of components vary over time: we continuously record the evaluation times of subqueries and adapt s_j .

3.4 Optimal Exploitation of Memory Resources

NN-search for large image databases using a single machine is IO-bound because the feature data is too large to fit into main memory, and CPU-costs are negligible in comparison. On the other hand, the total main memory of the components in the NOW may be large enough to hold the entire approximation data. To optimally exploit memory resources, each component should hold a partition of the approximation file in main memory, and the partition of the subquery of a component should be within this partition. Subsequently, we speak of *IO-bound NN-search* when referring to the first case, and *CPU-bound NN-search* otherwise.

It is desirable to eliminate the IO-boundness of NN-search. A strong advantage of the VA-File is that much less resources are needed to do so, as compared to the sequential scan which always works on the exact representation of the data points. Consider for example a feature set with 900 MB of data, and components with 100 MB of main memory. We need at least 9 machines to store the entire feature data in main memory. However, the VA-File only scans an approximation of the feature data. For instance, using 8 bits per dimension, the size of the approximation data is 4 times smaller than the one of the original data that uses 32 bits per dimension, i.e., only 225 MB. Consequently, the VA-File algorithm requires only 3 components to eliminate the IO-bottleneck.

Having discussed the benefit from good memory exploitation, what is the design alternative in this context? One might not want to reserve the main memory only for this task if there are other applications. We will investigate the benefit from the memory-intensive alternative in quantitative terms.

4 Performance Evaluation

We have conducted an extensive evaluation with feature data from a large image collection, as well as experiments with synthetic data. We have obtained

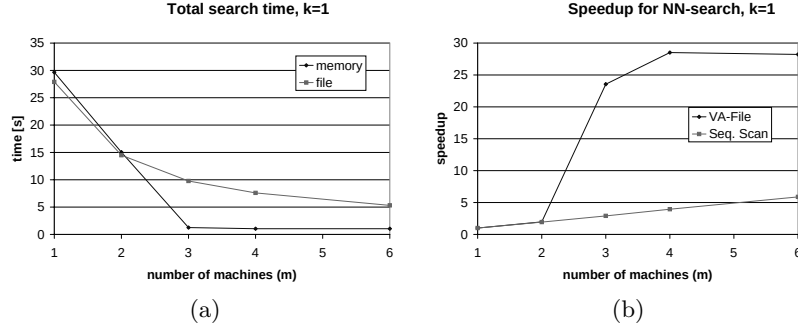


Fig. 3. Parallel VA-File-based search: (a) search times with and without optimal memory exploitation. (b) speedup compared to parallel sequential scan search.

the feature data by extracting several different features with a dimensionality ranging from 9 to 2240 from over 100,000 images. The amount of feature data has varied between 3 MB and 900 MB. Since our objective is to deal with large image collections, we limit the discussion to the largest feature representations with 900 MB of vector data and 230 MB of approximation data. We should point out that we have already studied the problem of updating data elsewhere [13, 17].

4.1 Optimal Memory Exploitation

The first set of experiments investigates the role of the main memory. Figure 3 (a) compares the search time of parallel VA-File for the two following techniques: (1) exploiting the memory available (denoted as *memory* in the figure), and (2) fetching data always from disk (denoted as *file*). The difference in the search times of the two techniques is not large when employing one or two machines. Namely, the approximation data (230 MB) does not fit into main memory. However, the effect becomes visible if the NOW consists of more than three machines. In this case, the search time of the *memory*-technique is less by a factor up to five. This is because disk IO is not necessary any more.

Figure 3 (b) demonstrates that the VA-File benefits more from this effect than the sequential scan. The figure compares the speedup of the response times with the VA-File with the one with sequential scan. Since the vector data is 900 MB, a NOW with 6 machines does not suffice to hold the entire vector data in main memory. On the other hand, the approximation data, only 230 MB, fits into main memory with more than three machines. Therefore, the speedup of the VA-File increases significantly between $m = 2$ and $m = 4$. In Figure 3 (b), this effect is not visible for the sequential scan. We could observe it if the NOW consisted of more than 9 machines. Further, notice that the speedup of NN-search is almost 25 when using three machines instead of one. Instead of 30 seconds

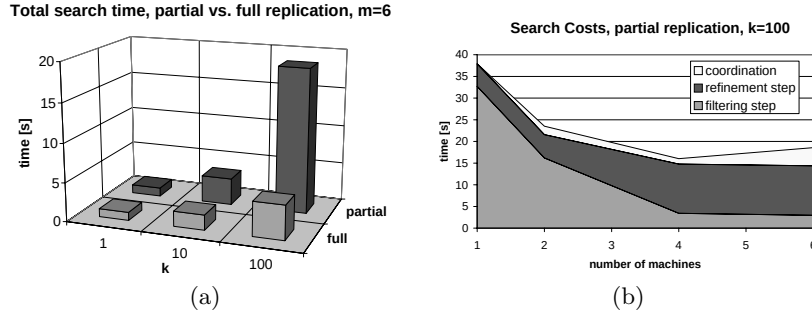


Fig. 4. (a) Total search time of partial and full replication with $m = 6$. (b) Search costs of partial replication in the case of $k = 100$ as a function of m .

for NN-search with one machine, VA-File-based search with three machines lasts only about one second for this particular case.

4.2 Partial vs. Full Replication of the Full Vector Data

With partial replication, access to the candidates in the refinement step is via networked file system to a central vector file. Figure 4 (a) compares the search times with partial replication with the ones with full replication. For this experiment, we have used 6 identical machines. The larger k becomes, the worse partial replication performs. This is because each component accesses at least k candidates in the central vector file. For instance, there are at least 600 accesses to the vector file for $m = 6$ and $k = 100$. Figure 4 (b) depicts this situation in more detail: it shows the coordination costs, the costs of the filtering step and the costs of the refinement step with partial replication as a function of m . k was set to 100. Obviously, the cost of the filtering step decreases with the number of machines in the NOW. On the other hand, the cost of coordination and the cost of the refinement step grow with m . This means that we do not really profit from the reduced filtering cost. We conclude that partial replication only pays off if $m \cdot k$ is small.

4.3 Heterogeneous Environments

In this series of experiments, we have used a heterogeneous NOW with 6 machines: 4 PentiumII machines, referred to as A , B , C , D , one AMD K6 machine (E), and one PentiumPro machine (F). We have compared the dynamic partitioning scheme (*individual*) with the alternative with equal partition sizes. Furthermore, we varied the number of subqueries for the later case. Figure 5 (a) plots the search costs of each component, as well as the total cost of search. Figure 5 (b) displays the number of data items searched by each component. The dynamic partitioning scheme yields the best overall search time. The search cost of each component is almost the same. Search performance with $n = m$

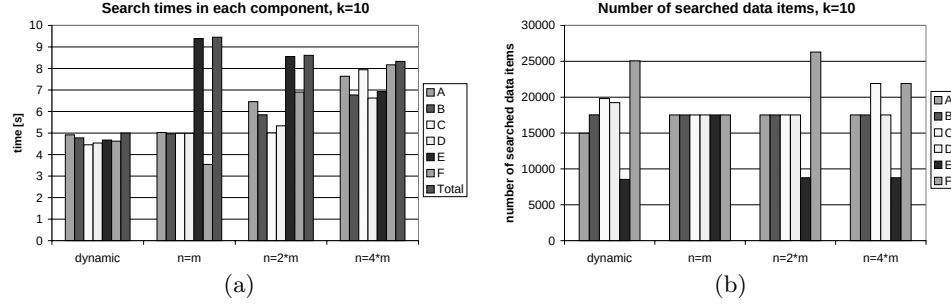


Fig. 5. A parallel, IO-bound VA-File based NN-Search: (a) search times of the components, (b) partition sizes of the components

is worst since machine E needs almost twice the time of machines A to D to evaluate its subquery. Increasing n results in a better overall search time since the faster machines evaluate more subqueries. Finally, the search time of each machine is about the same in the case of $n = 4 \cdot m$. But the overall search cost is more than 50% higher than with the dynamic load balancing technique due to the increased overhead.

5 Conclusions

NN-search is a common way to implement similarity search. But performance is a problem if dimensionality is high. We have argued that the preferred approach is as follows: one should use a linear algorithm and parallelize it and work with vector approximations. Our contribution is to identify the design alternatives, to relate them to the design objectives imposed by a Digital Library, and to evaluate the alternatives. Our implementation does reduce search time to a reasonable level for relatively large collections. The best speedup value observed with few resources, as compared to a monolithic system, is 30. We have demonstrated that our implementation allows for interactive-time similarity search, even over relatively large collections.

Interactive-time similarity search is particularly useful when the search consists of several steps. For instance, a relevance feedback engine, together with an intuitive user interface, is part of our DL architecture [3]. Preliminary experiments have shown that the relevance feedback functionality must go along with an efficient implementation in order to be accepted by the DL users.

This work is part of a larger effort to build a coordinated component-based DL architecture [17, 13]. This architecture allows to cope with the important requirements of autonomy, extensibility and performance, as [17, 13] explains. The last requirement is met because our architecture facilitates plug-and-play scalability of components. An online demonstration of the search capabilities of the system is available at <http://simulant.ethz.ch/Chariot/>.

References

1. S. Berchtold, C. Böhm, B. Braunmüller, D.A. Keim, and H.-P. Kriegel. Fast parallel similarity search in multimedia databases. In *Proc. of the ACM SIGMOD Int. Conf. on Management of Data*, pages 1–12, Tucson, USA, 1997.
2. S. Berchtold, D.A. Keim, and H.-P. Kriegel. The X-tree: An index structure for high-dimensional data. In *Proc. of the Int. Conference on Very Large Databases*, pages 28–39, 1996.
3. Jakob Bosshard. An open and powerful relevance feedback engine for content-based image-retrieval. Diploma thesis (in english), Institute of Information Systems, ETH, Zurich, 2000.
4. P. Ciaccia, M. Patella, and P. Zezula. M-tree: An efficient access method for similarity search in metric spaces. In *Proc. of the Int. Conference on Very Large Databases*, Greece, 1997.
5. P. Ciaccia, P. Tiberio, and P. Zezula. Declustering of key-based partitioned signature files. *ACM Transactions on Database Systems*, 21(3), September 1996.
6. Corbis. The place for pictures on the internet. <http://www.corbis.com/>.
7. M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content: The QBIC system. *Computer*, 28(9):23–32, September 1995.
8. A. Guttman. R-trees: A dynamic index structure for spatial searching. In *Proc. of the ACM SIGMOD Int. Conf. on Management of Data*, pages 47–57, Boston, MA, June 1984.
9. I. Kamel and C. Faloutsos. Parallel R-trees. Technical Report CS-TR-2820, University of Maryland Institute for Advanced Computer Studies Dept. of Computer Science, Univ. of Maryland, College Park, MD, January 6 1992.
10. NASA. Nasa's eos program. <http://www.eos.nasa.gov/>, <http://eosps.gsfc.nasa.gov/>, http://spsosun.gsfc.nasa.gov/New_EOSDIS.html.
11. G. Panagopoulos and C. Faloutsos. Bit-sliced signature files for very large text databases on a parallel machine architecture. *Lecture Notes in Computer Science*, 779, 1994.
12. A. N. Papadopoulos and Y. Manolopoulos. Similarity query processing using disk arrays. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 27(2), 1998.
13. H.-J. Schek and R. Weber. Higher-Order Databases and Multimedia Information. In *Proc. of the Swiss/Japan Seminar "Advances in Databases and Multimedia for the New Century – A Swiss/Japanese Perspective"*, Kyoto, Japan, December 1–2, 1999, Singapore, 2000. World Scientific Press.
14. Columbia University. Webseek: A content-based image and video search and catalog tool for the web. <http://disney.ctr.columbia.edu/webseek/>.
15. R. Weber and K. Böhm. Trading Quality for Time with Nearest-Neighbor Search. In *Advances in Database Technology — EDBT 2000, Proc. of the 7th Int. Conf. on Extending Database Technology, Konstanz, Germany, March 2000*, volume 1777 of *Lecture Notes in Computer Science*, pages 21–35, Berlin, 2000. Springer-Verlag.
16. R. Weber, H.-J. Schek, and S. Blott. A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In *Proc. of the Int. Conference on Very Large Databases*, New York, USA, August 1998.
17. Roger Weber, Jürg Bolliger, Thomas Gross, and Hans-J. Schek. Architecture of a networked image search and retrieval system. In *Eighth International Conference on Information and Knowledge Management (CIKM99)*, Kansas City, Missouri, USA, November 2–6 1999.

Dublin Core Metadata for Electronic Journals

Ann Apps and Ross MacIntyre

MIMAS, Manchester Computing, University of Manchester,
Oxford Road, Manchester, M13 9PL, UK
`ann.apps@man.ac.uk`, `ross.macintyre@man.ac.uk`

Abstract. This paper describes the design of an electronic journals application where the article header information is held as Dublin Core metadata. Current best practice in the use of Dublin Core for bibliographic data description is indicated where this differs from pragmatic decisions made when the application was designed. Using this working application as a case study to explore the specification of a metadata schema to describe bibliographic data indicates that the use of Dublin Core metadata is viable within the journals publishing sector, albeit with the addition of some local, domain-specific extensions.

Keywords. Dublin Core, metadata, bibliographic citation, electronic journals.

1 Introduction

Metadata is a description of an information resource, and hence can be thought of as ‘data about data’. Within the context of the World Wide Web metadata may be used for information discovery, but metadata is also important in the context of cataloguing resources. Dublin Core Metadata [1] is an emerging standard for simple resource description and for provision of interoperability between metadata systems.

The article header and abstract information used by an electronic journals application when publishing academic journal articles on the World Wide Web is effectively metadata for those articles. This paper describes an implementation of an electronic journals application, for a publisher, where the article metadata is held as Dublin Core, and some of the problems associated with specifying the metadata schema to develop the application design.

2 Dublin Core Metadata

The Dublin Core Metadata Element Set has been endorsed by the Dublin Core Metadata Initiative after its development by the Dublin Core Working Groups. These international working groups are open membership email discussion groups, with occasional ‘face-to-face’ meetings at the International Dublin Core Metadata Workshop series. The Dublin Core standard for metadata specification is primarily concerned with the semantics of the metadata rather than the

syntax used for its inclusion with an information resource. It is designed for simple resource description and to provide minimum interoperability between metadata systems, with a consequent potential for cross-domain metadata interchange. Dublin Core does not attempt to meet all the metadata requirements of all sectors, where Dublin Core would be enhanced to produce domain-specific metadata schemas for richer descriptions.

Basic Dublin Core (Version 1.1) is a fifteen metadata element set as developed by the Dublin Core Metadata Initiative: Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights. Detailed definitions of these elements are given on the Dublin Core Web Site [1]. All Dublin Core elements are optional and repeatable.

Qualified Dublin Core allows further refinement and specification of the element content using interoperability qualifiers to increase the semantic precision of metadata. The initial set of proposed qualifiers [2] has recently (April 2000) been approved by the Dublin Core Usage Committee, after many months of discussion by the element working groups. Qualified Dublin Core includes element refinement qualifiers such as ‘Alternative Title’ as well as the possibility of specifying encoding schemes for element values. The encoding schemes, which include controlled vocabularies and formal notations, may be international standards, for instance ISO639-2 for Language, or defined within Dublin Core.

2.1 Dublin Core in Practice

There have been problems associated with using Dublin Core in practical implementations. Firstly the basic element set may be insufficient to fully describe the resource within a particular domain. Secondly the timescale for Dublin Core metadata definition has meant that it has been necessary to anticipate the standard definitions before they have been fixed, sometimes incorrectly.

Basic Dublin Core Element Set. The basic Dublin Core element set does not attempt to meet all requirements of all resource domains. It is acknowledged that application designers will meet local functional requirements by the inclusion of additional elements and qualifiers within a local namespace. But it is expected that if local qualifiers are demonstrated to be of broader application they will attract wider deployment. Application designers will be encouraged to register their local qualifiers with the Dublin Core Metadata Registry [3] to leverage their usage, and hence interoperability, more widely.

Dublin Core Definition Timescale. Because of the nature of the method of defining Dublin Core through open-membership discussion groups, the length of time before versions have been baselined has been considerable. This is particularly the case with Qualified Dublin Core, whose first skeleton version has only recently been approved. This has caused a problem for application designers who have needed to produce an ‘up-and-running’ implementation quickly. Until now,

using Dublin Core has been rather like trying to ‘hit a moving target’. Inevitably some of the decisions made in these application designs have since been deprecated, but it is not always possible to change working applications easily.

3 An Electronic Journals Application

The Manchester University Press (MUP) [4] electronic journals application hosted by MIMAS [5], at the University of Manchester, UK, holds the article metadata as enhanced Dublin Core using an XML [6] syntax. Journal article metadata is supplied to MIMAS in SGML format, using the Simplified SGML for Serial Headers (SSSH) [7] Document Type Definition (DTD) for the majority of the journals but a typesetter’s proprietary DTD for one journal. During the data handling process this SGML metadata is transformed into the required ‘Dublin Core in XML’ format, and XML ‘Table of Contents’ files are generated, using specific OmniMark [8] programs written by MIMAS. The full article files are supplied as PDF and are provided in this format to the end-user. When an end-user requests viewing of the article metadata, including its abstract, the ‘Dublin Core in XML’ is transformed into HTML ‘on-the-fly’ using another OmniMark program written by MIMAS.

It was decided to hold the article header information in Dublin Core because it was recognised that this is essentially metadata for the articles. Thus exploration of using Dublin Core for bibliographic applications seemed appropriate. A common format for holding article metadata for all the journals was necessary because they are delivered to MIMAS by Manchester University Press in SGML using two different DTDs.

In general, the mapping between the article metadata elements and the Dublin Core Metadata Element Set was obvious. But decisions were needed how to capture: the journal article bibliographic citation information such as journal title, volume and issue numbers; the parts of the name of a Creator; the affiliation of a Creator; the location and size of the corresponding full article PDF file; the specific document Type; and the language of those elements which could be supplied in multiple languages. A pragmatic approach was taken to resolving these issues because timescale constraints obviated waiting for Dublin Core endorsed definitions. Details of how these were coded and of current best practice are described below.

Although all Dublin Core elements are optional, to implement a viable electronic journals application some elements must be mandatory. The application requires every article to have Title, Publisher, Date, Type, Format, Identifier, Source, Language, Relation, and Rights. This requirement is imposed by the XML DTD.

The design of this application was developed from work done at MIMAS on the Nature Digital Archive project [9], [10], and from previous work on the SuperJournal project [11]. Examples given within the paper are from a sample article header published in the January 2000 issue of the ‘International Journal of Electrical Engineering Education’ [12].

3.1 Simple Element Mappings

Some article metadata elements mapped obviously onto Dublin Core elements, as shown in Table 1.

Table 1. Mappings to Dublin Core elements.

Article Metadata	Dublin Core Element
Title	Title
Author	Creator
Keyword	Subject
Abstract	Description
Publisher name	Publisher
Cover date	Date
Article language	Language
Full article format	Format
Copyright	Rights

The Dublin Core elements Contributor and Coverage are not used. Multiple instances are allowed for Creator, and Subject, one for each keyword. Multiple instances are also provided for Title and Description to implement multiple language metadata.

4 Journal Article Citation Metadata

A major problem in using Dublin Core for journal article metadata is capturing the bibliographic citation information for an article within a journal issue. Since this electronic journals application was designed, this problem has been addressed by a Dublin Core Working Group who were specifically tasked with addressing this issue and recommending a solution. The decision made in this application has since been deprecated by the DC-Citation Working Group's [13] consensus, but the experience of designing the application was input to the working group of which MIMAS has active membership. Although a recommendation has been made by the DC-Citation Working Group it has not yet been endorsed by the Dublin Core Metadata Initiative and the requisite element qualifiers and encoding schemes have not yet been recommended by the relevant individual element working groups.

To capture a journal article citation for use within a bibliographic reference, the minimum requirement is the Journal Title, the Volume number and the Start Page number of the article within the printed journal. Generally bibliographic references also include the publication year. For article metadata within an electronic journals application it is also necessary to capture the number of the Issue containing the article. It is probably sensible to capture the ISSN number

of the journal and the End Page of the article within the printed journal. There are some defined schemes for specifying this information, in particular the Serial Item and Contribution Identifier (SICI) [14].

4.1 Journal Article Citation within MUP E-Journals

Within the Manchester University Press E-Journals application, the journal article citation information is held within Source, using structured values within the local namespace, ie. according to an 'internal' scheme, 'MUP.JNLCIT'. These structured values capture separately the Journal Title (JTL), the Volume (VID), the Issue number (IID), the Start Page (PPF) and the End Page (PPL). Holding this information as a structured value simplifies subsequent parsing and processing by other applications, such as the program which displays the article metadata to the end-user. Further citation information is held within Identifier, the SICI (Version 1) for the journal issue, and within Relation, the journal ISSN. For example, where the local namespace is 'MUP':

```
<MUP.Source SCHEME="MUP.JNLCIT">
  <MUP.JTL>International Journal of Electrical Engineering
    Education</MUP.JTL>
  <MUP.VID>37</MUP.VID>
  <MUP.IID>1</MUP.IID>
  <MUP.PPF>26</MUP.PPF>
  <MUP.PPL>37</MUP.PPL>
</MUP.Source>
<DC.Identifier SCHEME="SICI">0020-7209(20000101)37:1</DC.Identifier>
<DC.Relation SCHEME="ISSN" RELATION="IsPartOf">0020-7209</DC.Relation>
```

4.2 DC-Citation Recommendation

The final recommendation of the DC-Citation Working Group was to hold the journal article bibliographic citation information, including page range, within Identifier, using either a structured value or a prescribed syntax within the text string content of the Identifier value. Using structured values would make subsequent parsing and processing of the Identifier simpler, but these have not yet been ratified by the Dublin Core Metadata Initiative. If agreed by the Identifier Working Group, this use of Identifier would be qualified with a 'Citation' qualifier. Because all Dublin Core elements are repeatable, it would be possible to have additional instances of Identifier containing the journal article identifier encoded according to other schemes such as SICI or Digital Object Identifier (DOI) [15]. Relation will hold the citation information for the next level above. So for an article the Relation 'IsPartOf' would indicate the bibliographic citation for the journal issue. An article could additionally be identified using another encoding scheme such as SICI or DOI. Its containing issue could be indicated using DC.Relation 'IsPartOf'.

Using this recommendation and a structured value 'DCCITE', the above example could be encoded as:


```

<DC.Identifier.Citation SCHEME="DCCITE">
  <JournalTitleFull>International Journal of Electrical Engineering
    Education</JournalTitleFull>
  <JournalTitleAbbreviated>IIEEE</JournalTitleAbbreviated>
  <JournalChronology>January 2000</JournalChronology>
  <JournalVolume>37</JournalVolume>
  <JournalIssueNumber>1</JournalIssueNumber>
  <JournalPages>26-37</JournalPages>
</DC.Identifier.Citation>
<DC.Identifier SCHEME="DOI">10.1060/IIEEE.2000.003</DC.Identifier>
<DC.Relation.IsPartOf SCHEME="SICI">
  0020-7209(20000101)37:1<DC.Relation.IsPartOf>

```

‘JournalChronology’ is included in this recommended journal article citation metadata to overcome issues surrounding the use and semantics of DC.Date to indicate the journal cover date, which is essentially an artificial date but necessary for journal cataloguing and discovery. An additional advantage to including JournalChronology within the citation is the ability to capture dates such as ‘Spring 2000’ which appear on some journal issues but are difficult to encode.

If a structured value were not used for ‘DCCITE’, for instance within an HTML ‘META’ tag, the citation would be specified as a semi-colon separated text string according to a defined syntax.

There was much discussion within the DC-Citation Working Group before this recommendation was made following the 7th International Dublin Core Workshop in Frankfurt, Germany in October 1999. As well as the above coding method in the MUP E-Journals application using Source, it was suggested that the journal article citation information could be in Relation using the ‘IsPartOf’ qualifier.

Although these recommendations have been made by the DC-Citation Working Group, the indicated qualifiers and encoding schemes must be recommended by the relevant Dublin Core Element Working Groups and ratified by the Dublin Core Metadata Initiative. Currently the only approved encoding scheme for Identifier is ‘URI’ and it has no approved element qualifiers.

5 Author Name and Affiliation

Although ‘Author’ maps obviously onto DC.Creator, there are currently no approved qualifiers or encoding schemes for DC.Creator. The element value is simply a Creator’s name as a free-text string. Within an electronic journals application it is desirable to split the author’s name into constituent parts. For instance, indexing on authors’ surnames could provide useful functionality to end-users. Within the MUP E-Journals application, the author names are encoded using a local structured value, ‘MUP.AU’ which captures separately an author’s family name, first names and an optional suffix.

A further problem is to capture an author’s affiliation. For an article published in an academic journal this affiliation indicates the author’s institution at

the time when the article was published, which is not necessarily the author's current address. It would have been possible to extend the above creator structured value, 'MUP.AU', to additionally capture this affiliation. But if more than one author has the same affiliation, this would be repeated for each author. Some authors have more than one affiliation if they are associated with more than one institution. As well as not wishing to repeat addresses within the information displayed to the end-user, it seemed better to replicate the information supplied in the original typeset SGML where an address is defined once with pointers to it from the author names. Thus a further local scheme, 'MUP.AFFS' is defined to capture addresses including an identifier attribute. In addition, Creator has a local attribute pointing to the relevant address identifiers, which may be a comma-separated list.

Using as an example the same article as in the previous examples, the author details are encoded as:

```
<MUP.Creator SCHEME="MUP.AU" IDS="AFF1">
  <MUP.FNMS>Bill</MUP.FNMS><MUP.SNM>Olivier</MUP.SNM>
</MUP.Creator>
<MUP.Creator SCHEME="MUP.AU" IDS="AFF1">
  <MUP.FNMS>Oleg</MUP.FNMS><MUP.SNM>Liber</MUP.SNM>
</MUP.Creator>
<MUP.Creator SCHEME="MUP.AU" IDS="AFF2">
  <MUP.FNMS>Paul</MUP.FNMS><MUP.SNM>Lefrere</MUP.SNM>
</MUP.Creator>
<MUP.Creator SCHEME="MUP.AFFS">
  <MUP.AFF ID="AFF1">University of Wales</MUP.AFF>
  <MUP.AFF ID="AFF2">The Open University</MUP.AFF>
</MUP.Creator>
```

This specification scheme for author information does not allow for any grouping of authors. Author grouping is used in article header metadata by some publishers to indicate those authors whose affiliation is the same. It may also be used to indicate significant groupings of those contributing to an article. The journals included in this application do not make use of author grouping, although the SGML DTDs used for data supply would allow it, so this was not a requirement within the described application. But author grouping could have been employed as an alternative solution to the problem of avoiding address repetition. Dublin Core does not include any notion of grouping of the elements, which are all optional and repeatable, but it would be possible to impose grouping by the syntax used in an actual implementation.

6 Other Issues

6.1 Full Article Format and File Size

It is necessary to capture in some way the format of the full text article to which the article metadata refers. Although within this application all articles are in

PDF format, other electronic journal applications may offer full text articles in a choice of formats, such as HTML in addition to PDF. The full article format is captured within DC.Format. It is additionally required to hold the size of the PDF file so that an end-user could be informed for download. The internal path and name of the file is useful to the application, though the path is possibly deducible using internal naming conventions. Full article file path and size are encoded using a local qualifier to DC.Relation, 'IsAbstractOf', with the PDF file size as an additional attribute. For example:

```
<DC.Format SCHEME="IMT">application/pdf</DC.Format>
<MUP.Relation SCHEME="MUP.PDF" RELATION="IsAbstractOf"
  PDFSIZE="99">IJEEE/V37I1/370026.pdf</MUP.Relation>
```

Using the now approved Dublin Core Format element qualifiers, and Identifier with a local encoding scheme, this would be better coded as:

```
<DC.Format.Medium SCHEME="IMT">application/pdf</DC.Format.Medium>
<DC.Format.Extent>99</DC.Format.Extent>
<MUP.Identifier SCHEME="MUP.PDF">
  IJEEE/V37I1/370026.pdf</MUP.Identifier>
```

6.2 Journal Article Type

The currently defined encoding scheme for DC.Type, the DCT1 Type Vocabulary, does not provide any means of indicating that the metadata is for a journal article. The 'DCT1' encoding scheme is at a higher level of abstraction. At present, its approved terms are: Interactive Resource; Dataset; Event; Image; Sound; Service; Software; Collection; Text. Using this scheme the article would simply be indicated as text. In addition to research articles an electronic journals application will contain 'Table of Contents' files, and could contain other types of document. Thus a local encoding scheme, 'MUP.TYPE' is also used.

```
<DC.Type SCHEME="DCT1">Text</DC.Type>
<MUP.Type SCHEME="MUP.TYPE">Research Article</MUP.Type>
```

A list of document types necessary for an electronic journals application may include: Announcement; Book Review; Corrigendum; Critique; Editorial; Erratum; Discussion Forum; Invited Commentary; Letter to the Editor; Obituary; Personal View; Research Note; Research Article; Review Article; Short Communication; Special Report; Table of Contents; Technical Report. Possibly in the future a list of types suitable for the academic journals publishing sector will be registered as a domain-specific controlled vocabulary with Dublin Core.

6.3 Language

Although all of the articles within this electronic journals application are written in English, one of the Manchester University Press journals has article titles

and abstracts additionally in French, German and Spanish. These have to be captured within the article metadata, and are displayed to the end-user when viewing article information. Thus Title and Description have a language attribute and may have multiple instances. An example encoding for Title follows, the encoding for Description being similar.

```
<DC.Title LANGUAGE="en">Specifications and standards for learning
    technologies: the IMS project</DC.Title>
<DC.Title LANGUAGE="fr">Spécifications et normes pour
    technologies de formation: le projet IMS</DC.Title>
<DC.Title LANGUAGE="de">Spezifikationen und Normen für
    Lerntechnologien: das IMS Projekt</DC.Title>
<DC.Title LANGUAGE="es">Especificaciones y estándares para las
    tecnologías de la enseñanza: el proyecto
    IMS</DC.Title>
```

Non-keyboard characters are held as SGML character entities within the article metadata, for example ‘é’ is encoded as ‘é’ and ü as ‘ü’. Most of the characters used in European languages are displayed correctly by Web browsers using SGML character entity encoding. Other characters are converted to a displayable encoding when converted to HTML for end-user display of an article’s metadata, using the SGML-aware language, OmniMark.

7 Conclusion

The experience of designing and implementing this electronic journals application has provided a case study to explore the viability of using Dublin Core for journal article metadata. It has indicated that it is possible to use Dublin Core within the academic journals publishing sector with the addition of some local encoding schemes, in particular structured values. It was not necessary to include new local metadata elements. Any future new design of an electronic journals application would take into account more recent decisions about Qualified Dublin Core, in particular the recommendations made by the Dublin Core Citation Working Group.

There are initiatives to progress Dublin Core into an international standard for metadata for simple resource description and minimum interoperability. MIMAS is a member of the European CEN/ISSS Workshop on Metadata for Multimedia Information – Dublin Core (MMI-DC) [16] which seeks to ratify the use of Dublin Core metadata as a standard within Europe as well as providing guidelines on its use and an Observatory on European projects which are using Dublin Core.

Using a standard metadata description such as Dublin Core for an application will assist in future interoperability with other applications. An application may wish to provide cross-domain searching, for instance across both research datasets and the corresponding research literature. Or it may wish to display only a subset of the information, for instance on a mobile phone screen.

Up until now, the main problem of using Dublin Core metadata to implement a production quality application has been the lack of a stable definition. But Basic Dublin Core Version 1.1 is now fixed, and the first version of Qualified Dublin Core has recently been announced to Dublin Core Working Group members. The Dublin Core Metadata Initiative intend to set up a Dublin Core Metadata Registry for the registration of local, domain-specific elements, qualifiers and encoding schemes in addition to those endorsed by Dublin Core. This will allow sharing of interoperable metadata items and leverage the approval of further qualifiers and encoding schemes by the Dublin Core Metadata Initiative for general use. So it appears that Dublin Core metadata is now mature enough, and acceptable as a standard, for its use to be recommended wherever metadata is required.

References

1. Dublin Core Metadata web site. <http://purl.org/dc/>
2. Weibel, S.: Approval of Initial Dublin Core Interoperability Qualifiers. Email message to DC-General Working Group (2000) <http://www.mailbase.ac.uk/lists/dc-general/2000-04/0010.html>
3. Dublin Core Metadata Registry Working Group. <http://www.mailbase.ac.uk/lists/dc-registry/>
4. Manchester University Press web site. <http://www.manchesteruniversitypress.co.uk>
5. Electronic Publishing at MIMAS web site. <http://epub.mimas.ac.uk/>
6. XML. <http://www.w3.org/XML>
7. SSSH, Simplified SGML for Serial Headers, DTD. <http://www.oasis-open.org/cover/gen-apps.html#sssh>
8. OmniMark Technologies web site. <http://www.omnimark.com>
9. MacIntyre, R., Tanner, S.: Nature - a Prototype Digital Archive. International Journal on Digital Libraries (2000), Springer-Verlag, (Scheduled for 2000(2))
10. Apps, A., MacIntyre, R.: Metadata for the Nature Digital Archive. <http://epub.mimas.ac.uk/natpaper.html>
11. Apps, A.: SuperJournal Metadata Specification. SuperJournal Project Report, <http://www.superjournal.ac.uk/sj/sjmc141.htm>
12. Olivier, B., Liber, O., Lefrere, P.: Specifications and standards for learning technologies: the IMS Project. International Journal of Electrical Engineering Education **37**(1) (2000) 26-37 doi://10.1060/IJEEE.2000.003 (<http://dx.doi.org/10/1060/IJEEE.2000.003>)
13. Dublin Core Bibliographic Citations and Versions Working Group. <http://purl.org/dc/groups/citation.htm>
14. SICI, Serial Item and Contribution Identifier. <http://sunsite.berkeley.edu/SICI/>
15. DOI, Digital Object Identifier. <http://www.doi.org>
16. CEN/ISSS Workshop on Metadata for Multimedia Information (MMI-DC). <http://www.cenorm.be/iss/Workshop/MMI-DC>

An Event-Aware Model for Metadata Interoperability

Carl Lagoze¹, Jane Hunter², and Dan Brickley³

¹Cornell University, Ithaca, NY, USA
lagoze@cs.cornell.edu

²DSTC Pty. Ltd., Brisbane Australia
jane@dstc.edu.au

³ILRT, Bristol, UK
daniel.brickley@bristol.ac.uk

Abstract. We describe the ABC modeling work of the Harmony Project. The ABC model provides a foundation for understanding interoperability of individual metadata modules – as described in the Warwick Framework – and for developing mechanisms to translate among them. Of particular interest in this model is an *event*, which facilitates understanding of the lifecycle of resources and the association of metadata descriptions with points in this lifecycle.

1. Metadata Modularity and Interoperability

The Warwick Framework [23] describes the concept of modular metadata - individual metadata packages created and maintained by separate communities of expertise. A fundamental motivation for this modularity is to scope individual metadata efforts and encourage them to avoid attempts at developing a universal vocabulary. Instead, individual metadata efforts should concentrate on classifying and expressing semantics tailored toward focused functional and community needs. Warwick Framework-like modularity underlies the design of the W3C's Resource Description Framework (RDF) [15, 24], which is a modeling framework for the integration of diverse application and community-specific metadata vocabularies.

An outstanding challenge of such modularity is the interoperability of multiple metadata packages that may be associated with and across resources. Metadata packages are by nature not semantically distinct, but overlap and relate to each other in numerous ways. Achieving interoperability between these packages via one-to-one crosswalks [4] is useful, but this approach does not scale to the many metadata vocabularies that will continue to develop. A more scalable solution is to exploit the fact that many entities and relationships - for example, people, places, creations, organizations, events, and the like - are so frequently encountered that they do not fall clearly into the domain of any particular metadata vocabulary but apply across all of them.

The Harmony Project [6] is investigating this more general approach towards metadata interoperability and, in particular, its application in multimedia digital libraries. This approach, the *ABC model and vocabulary*, is an attempt to:

- formally define common entities and relationships underlying multiple metadata vocabularies;
- describe them (and their inter-relationships) in a simple logical model;
- provide the framework for extending these common semantics to domain and application-specific metadata vocabularies.

The concepts and inter-relationships modeled in ABC could be used in a number of ways. In particular, individual metadata communities could use these underlying concepts (the ABC model) to guide the development of community-specific vocabularies. These individual communities could use formalisms such as RDF to express the possibly complex relationships between the ABC model and their community-specific vocabularies. Furthermore, the formal expression of the relationships between community-specific vocabularies and the ABC model could provide the basis for a more scalable approach to interoperability among multiple metadata sets. Rather than one-to-one mappings among metadata vocabulary semantics, a more scalable basis for interoperability could be achieved by mapping through this common logical model.

This paper describes the initial results of our work on the ABC model, the main focus of which is describing *events* and their role in metadata descriptions. Briefly stated, our argument is as follows. Understanding the relationship among multiple metadata descriptions (and ultimately the vocabularies on which they are based) begins by understanding the entities (resources) they purport to describe. Understanding these entities entails a comprehension of their lifecycle and the events, and corresponding transitions and transformations, that make up this lifecycle.

This work is influenced by and builds on a number of foundations. The significance of events and processes in understanding knowledge has deep routes in philosophy [12]. The importance of processes and events in resource descriptions has been recognized by a number of communities including the bibliographic community [5], museums [21], the archival community [11], and those concerned with e-commerce and rights management [7]. Events, and their role in metadata interoperability, were recognized in [10]. Our modeling principles are influenced by work in the W3C and related communities, where both the XML Schema [13, 28] and RDF Schema [15] initiatives are evolving with the goal of formally modeling and representing data (and metadata) on the Web. These efforts and our own build on work in the database community to understand, model, and query semi-structured data [9].

The remainder of this paper is structured as follows. Section 2 describes the relevance of events for understanding the relationship between metadata vocabularies. Section 3 situates the concept of events and resource relationships within the broader ABC logical model. Section 4 then presents a formal model of events using UML [14] and then expresses this model using the XML Schema language. Section 5 uses this model to describe a compound multimedia example. The paper closes with Section 6 that describes future directions.

2. Event-Aware Metadata

In January, 2000 the Harmony Project sponsored a workshop [1] that brought together representatives of several metadata initiatives to discuss interoperability. Subsequent to establishing shared perspectives and goals, the workshop focused on the importance of events in understanding intellectual resources, the nature of various descriptions of them, and the relationships between these descriptions. There was general agreement that a model to facilitate mapping between metadata vocabularies needs to be *event-aware*. This requirement builds on a number of observations about resources and descriptions. These observations are as follows.

As described in the IFLA FRBR (Functional Requirements for Bibliographic Records) [5], intellectual content evolves over time. The taxonomy developed in FRBR is a useful foundation for understanding the lifecycle of a single resource: it begins as a conceptual *work*, it may evolve into one or more *expressions* (e.g., an opera, a story, a ballet), these expressions may be realized in one or more *manifestations* (e.g. an edition or printing of a story in book form), and eventually these manifestations are disseminated as individual items (e.g., an individual copy of a book). The FRBR model largely applies to the evolution of a single resource; the subtleties of inter-resource relationships and the derivative nature of relationships between them also need to be understood.

An important aid towards understanding this evolution of an individual resource and the derivative relationships between resources is to characterize the events that are implicit in the evolution or derivation. For example, the evolution from *work* to *expression* may contain an implicit *composing event*. The process of making implicit events explicit – making them *first-class objects* – may then provide attachment points for common descriptive concepts such as agency, dates, times, and roles. A model that explicitly represents the attachment of these concepts to events may be useful for mapping between metadata vocabularies that express these concepts.

Events are also important in understanding metadata descriptions because of the way that they transform "input" resources into "output" resources, and the respective descriptions (or metadata) for those input and output resources. In particular, an event is important from a certain descriptive community's perspective because of the way the event changes a property of a resource that is of interest to that community. While an event changes one or more properties of a resource, other properties remain unchanged. For example, a "translation event" of *War and Peace* may change its language from Russian to English, but its author is still Fyodor Dostoyevsky.

Descriptive communities can be distinguished by the events that are of significance to them. For example, a community that focuses on the history of production of a film may consider the "event" associated with the insertion of a certain scene into a film significant. As a result that event may be explicit in their descriptive vocabulary – for example, that community may have a metadata attribute that describes the date of the scene insertion. Another community, say one concerned with the presentation of that film on a screen, may consider that event irrelevant and may consider the "is part of" relationship of the scene to the movie completely non-event related.

A particular metadata description is often a portrayal of a *snapshot* of some entity taken in a particular state - a perceived stability of the entity over a particular time and place that perforce elides events or lifecycle changes that are outside the domain of

interest by the particular descriptive community. The granularity of that snapshot (and the number of elided or revealed events) varies across metadata vocabularies. For example, a Dublin Core description [3], intended for relatively basic resource discovery, is a particularly coarse granularity snapshot. A Dublin Core description of a postcard of the Mona Lisa might list Leonardo Da Vinci as the creator even though numerous events took place on the portrayal of the Mona Lisa since the depiction by Da Vinci. On the other hand, an INDECS [7] description, for which the events associated with transfers of rights are extremely important, might describe more fine-grained event snapshots.

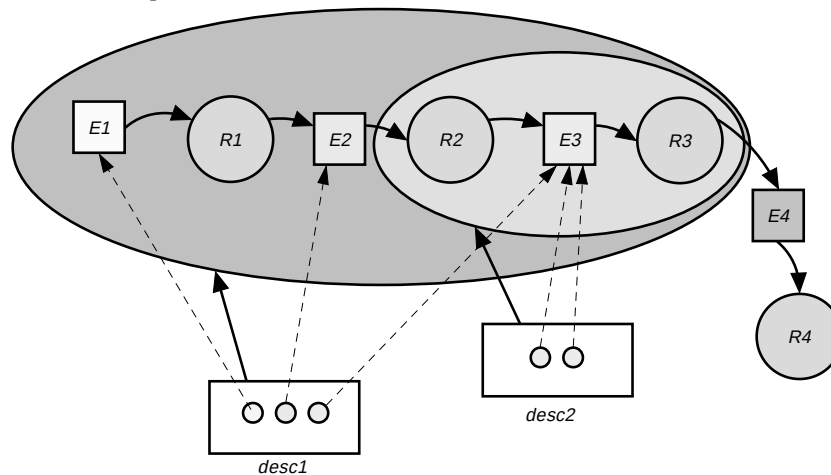


Fig. 1. Metadata and events

These observations suggest the following intellectual, and ultimately, mechanical approach towards understanding the relationships between metadata vocabularies:

- Develop a consistent and extensible model for events. This is the main subject of the remainder of this paper.
- Analyze the nature of the snapshots underlying the descriptions. For example, a coarse granularity Dublin Core description of a resource may combine attributes that span a number of transitions in the lifecycle of the resource. An INDECS description of what may at first seem like the “same” resource may actually focus on a smaller snapshot such as the attributes associated with a single transfer of rights in a contractual transaction.
- Attempt to interpolate and model the events that are contained within these snapshots and model these event transitions. For example, a single Dublin Core record may contain information about an agent who is a creator, an agent who is a translator, and an agent who is a publisher. This implies that the DC record actually describes a snapshot that implicitly contains three events: creation, translation, and publishing. Modeling these events would then permit the explicit linkage between the attributes of the respective description with the corresponding

event (e.g., associating the "Creator Agent" with the "agent event and associating the "Creation Date" of the description with the same event).

- Examine the overlap between the snapshots described by the individual descriptions. For example, the set of events implicit within an INDECS description may be fully contained within the broader snapshot of events within a DC description.
- Examine the relationship of the events in the event-aware models of the individual descriptions and of the properties that are associated with those events. Such event-aware analysis may make it possible to establish the relationship between the vocabulary-specific properties that "map down" to these events.

These concepts are illustrated in **Figure 1**. The larger circles represent manifestations of a resource as it moves through a set of event transitions; the events are represented by the squares interspersed between the circles. For example, event *E1* may be a creation event that produces resource *R1*. This resource may then be acted on by a translation event - event *E2* - producing resource *R2* and so on. The rectangles at the bottom of the figure represent metadata descriptions (instances of particular metadata vocabularies), and the ellipses that enclose part of the resource/event lifecycle represent the snapshot of the lifecycle addressed by that particular metadata description. For example, the larger dark-shaded ellipse represents the snapshot described by *desc1*, and the smaller light-shaded ellipse the snapshot described by *desc2*. The smaller circles within each descriptive record are the actual elements, or attributes, of the description. The dotted lines (and the color of each circle) indicate the linkage of the metadata element to an event - as shown the elements in *desc1* are actually associated with three different events that are implicit in the snapshot. For example, the attributes (moving from left to right) may describe *creator*, *translator*, and *publisher*, which are actually "agents" of the events. As shown, the three rose colored elements are all associated with a single event *E3*, implying a relationship between them that can be exploited in mapping between the two descriptive vocabularies that form the basis for the different descriptions.

3. The ABC Logical Model for Metadata Interoperability

The ABC logical model is built on a number of fundamental concepts and assumptions including universally identified resources, properties (as a special type of resource), and classes that create sets of resources (and properties). The model also defines a set of fundamental classes (sets of resources) including creations, events, agents, and relationships. These fundamental classes provide the building blocks for expression (through sub-classing) of application-specific or domain-specific metadata vocabularies. The reader is referred to [16] for more details on the complete model.

Of particular applicability to this paper is the multiple-view modeling philosophy in ABC. This allows properties (relations between resources) to be expressed in a simple binary manner or in a more complex manner that promotes the relation to a first-class resource. These first-class resources then provide the locus for associating properties that describe the relation.

Resources are related in numerous ways: containment, translation, and derivation are but three of the more common relations. Describing these relations is an important aspect of metadata. In some vocabularies (e.g., Dublin Core) these relation descriptions are rather simple; in others there is the need for increased descriptive power. ABC (by adopting RDF's graph data model) allows us to move between simple and complex relation descriptions as follows. We create a model in which the entity that is the input to the relation, the entity that is the output of the relation, and the relationship between the two entities are all represented as *resources*. In order to more richly describe these resource, we can then associate properties with them. In this manner, we have promoted - "reified" - a simple *relationship arc* to a first class resource and associated properties with it. For certain applications, the complicated, explicit model is most useful; other times it is better to have a simple, flattened representation of the 'real' state of affairs. In both cases it is useful to understand how the two representations inter-relate.

The example in Figure 2 illustrates this point with the "hasTranslation" relation. We can take a simple view and say just that some document has a translation into another document.

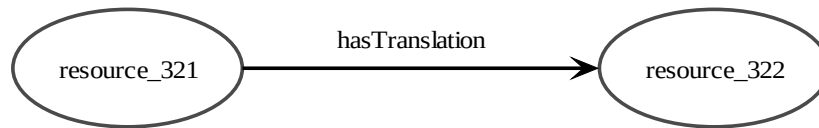


Fig. 2. Simple Resource Relationship

An alternative is shown in Figure 3 where we take a complex view and promote the *hasTranslation* relationship to a first class event resource. We can associate properties with that event to describe its details, such as its agents and its inputs and outputs. These details are the subject of Section 4.

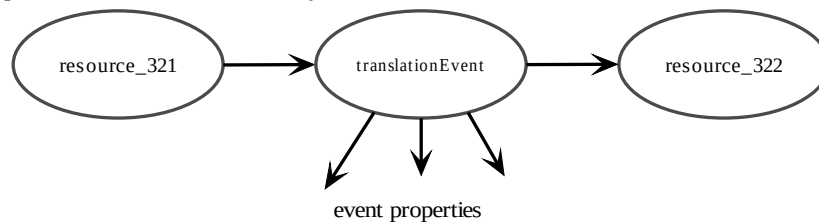


Fig. 3. Promoting Relationship to a First-Class Resource

This approach is applicable to a cross-section of events that have input and output resources or that describe an agent's contribution to a resource. Examples of such event/relation pairings include...

- Modification event <=> *VersionOf* relation
- Compilation event <=> *CompiledFrom* relation;
- Extraction event <=> *ExtractedFrom* relation;
- Reformat event <=> *IsFormatOf* relation;

- Translation event $\Leftrightarrow$ *TranslationOf* relation;
- Derivation event $\Leftrightarrow$ *DerivedFrom* relation.

In such cases, ABC provides two representational options and recipes for inter-conversion. When rich information is required, ABC provides the event model. This involves describing the *event* through which that relationship was realised as an object in itself, describing the hidden detail implicit in a simple binary relation. When concise/simple metadata is needed, flatter relations are used.

4. Modeling Events Using the ABC Vocabulary

The goal of the ABC vocabulary is to define and declare a core set of abstract base classes that are common across metadata communities. These base classes are intended to provide the attachment points for different properties (or metadata) that are associated with information content and its lifecycle. They will provide the fundamental infrastructure for modeling metadata and for refinement through subclassing.

A review of a number of metadata models (including IFLA [5], CIDOC [21], INDECS [27], MPEG-7 [8], and Dublin Core[3]) reveals the following common entities:

- Resources
- Events
- Inputs and Outputs
- Acts (and associated Acts and Roles)
- Context (consisting of Time and Place)
- Event Relations

A UML model of these entities and their relationship to each other is shown in the UML [26] model illustrated in Figure 4. This model can be represented declaratively in a schema definition using XML DTDs, RDF Schema [15] or XML Schema Language [13, 28]. The full version of this paper [22] provides an XML Schema representation.

The remainder of this section describes the entities in this model and our basic approach to an underlying metadata modeling framework. This approach is not final and will continue to be refined through implementation and feedback.

4.1. Resources

These represent the superclass of all of the possible things within our universe of discourse - they may be physical, digital or abstract. Every resource has a corresponding unique identifier.

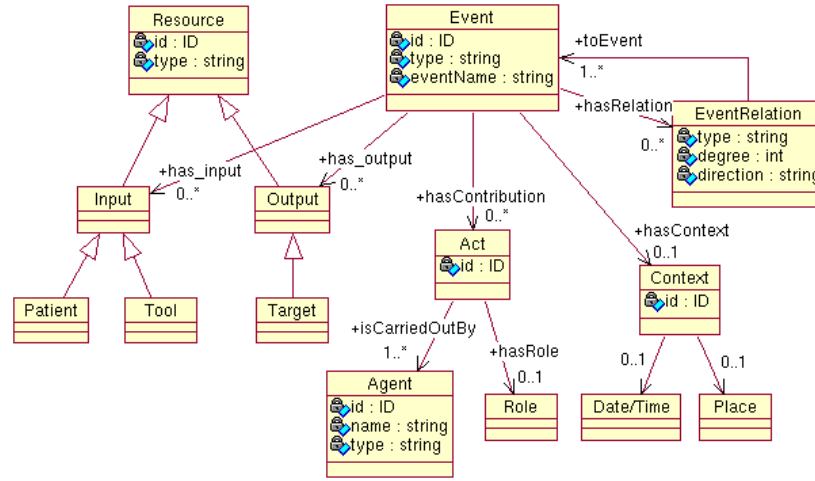


Fig. 4. UML representation of the basic event model

4.2. Events

An event is an action or occurrence. Every event has a Context (Time and/or Place) associated with it (although it may not always be explicit). Events may also have inputs and/or outputs associated with them. For example, events which generate a new or transformed resource (e.g. translation, modification) will have both input(s) and output(s).

The event class has the following properties:

- An eventType property (which may be enumerated);
- An optional eventName property;
- Optional input and output resources;
- Zero or more Act properties - which describe the contributions made by various agencies to the event;
- Zero or more EventRelations. These are relationships with other events and include relations such as the *contains* relation to define subEvents.

4.3. Inputs and Outputs

Events can have *Input* resources and/or *Output* resources.

Input resources vary in that some inputs are actually operated on during the event (Patients) whilst others are simply tools or references which are used during the event (Tools). The *Patient* and *Tool* subclasses of Input have been provided to support this distinction. This is important to avoid ambiguity during the complex-to-simple transformation when there are multiple inputs. Sometimes it may be difficult to

determine when a resource should be defined as an Input Tool and when it should be defined as an Agent. If there is a need to define the *Role* of the Input resource, then it must be defined as an Agent class.

Output resources vary in that some resources are the primary target outputs whilst others (e.g. messages) are of secondary importance. This distinction is important during metadata simplification in order to determine which Inputs and Acts are associated with which Output resources. The *Target* subclass is provided to prevent ambiguity and clearly specify the target output resources. Target output resources are assigned the Role/Agent property/value pairs during the complex-to-simple transformation.

4.4. Acts

An Act is a contribution to an event which is carried out by one or more actors or agents playing particular roles. An Act can only exist as a property of an Event. Each Act has one or more Agent properties and an optional Role property.

4.5. Agents

Agents represent the resources which act in an event - or the "actors" in an event. *Agents* are properties of *Acts* and usually have (through those Acts) an associated *Role* which defines the role that this actor plays in the particular event. The precise model by which agent roles are described is an area of ongoing research within Harmony. Some commonly-used agent types are:

- person/human being;
- organisation;
- instrument (hardware, software, machine).

In reality, any resource may take a causative role, thus allowing it to act as an agent. Additional possible agent types include: animals, fictional animals (Teletubbies), aliens, supernatural beings, imaginary creatures, inanimate objects (e.g., a painting that falls from a wall and strikes a sculpture, which shatters and then is presented as a new resource in a museum show), natural or environmental processes (storms, plagues, erosion, decay etc.).

4.6. Context

Date/Time. Time can be specified in a variety of ways. It can be either free text describing a period or event or a specific date/time format. It may also be either an instantaneous time or a time span. It may be GMT, local time or a time relative to a particular object's scope e.g. a time stamp in a video. Some examples include:

- The Battle of Hastings
- Next Year
- 21-10-99
- 00:07:14;09 - 00:12:36;21

Place. The place entity describes a spatial location. It can be free text or formatted. It can be absolute or relative. It can be a point, line, 2D or 3D region. Similarly to time, place can vary enormously in granularity. It may be a real world spatial location or a spatial location relative to a particular origin, coordinate system or objects' dimensions. Some examples of valid place values are shown below:

- 24 Whynot St, West End
- 0, 0, 100, 100
- Mars
- latitude, longitude
- the bottom left hand corner i.e. a section of a digital or physical object

4.7. EventRelations

EventRelations are provided to express relationships between Events. Typical top-level subtypes of EventRelations include: temporal, spatial, spatio-temporal, causal, conditional. Each of these may have enumerated subtypes e.g. temporal relations may include: *precedes*, *meets*, *overlaps*, *equals*, *contains*, *follows*. EventRelations may also have direction (*uni-directional*, *bi-directional*) and degree (unary, binary, n-ary) attributes associated with them. Conditional relations will have one or more condition statements associated with them.

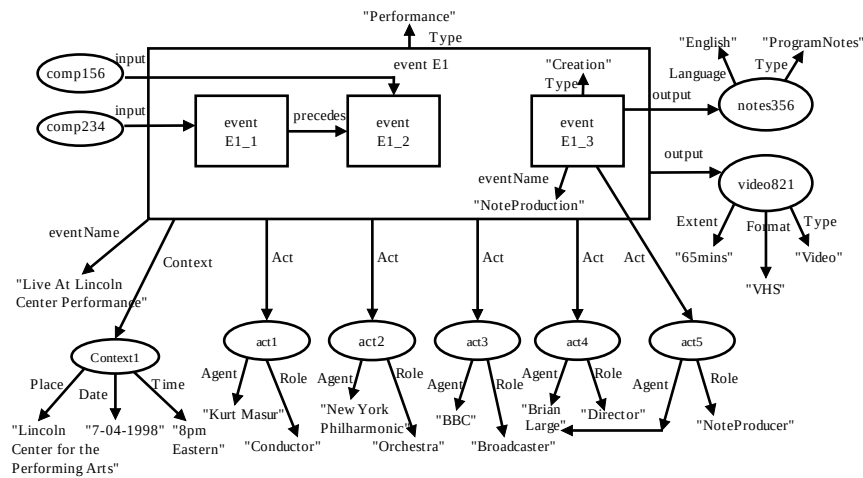


Fig. 5. Event-aware Model of A Performance of 'Concerto for Violin'

5. Applying the Model to a Complex Object

The following example of a complex object was developed at the January 2000 Harmony workshop.

A 65 min video (VHS) of a "Live at Lincoln Center Performance". The conductor is Kurt Masur. The Orchestra is the New York Philharmonic. The performance was

on April 7, 1998 at 8PM Eastern Time. The performance was broadcast live and recorded by the BBC. The direction and program notes (in English) were by Brian Large. The two pieces performed are:

- The Rite of Spring by Igor Stravinsky, written in 1911. Its length is 35 minutes
- Concerto for Violin by Phillip Glass written in 1992. With Robert McDuffie solo on the Violin. Its length is 25 minutes.

Figure 5 is an RDF model of the scenario based on the ABC vocabulary. The performance actually consists of 3 parts or sub-events, *event1_1*, *event1_2*, *event1_3*. *event1_1* and *event1_2* are the sequential performance sub-parts which are expressions of separate *concepts* or *works*. Figure 5 illustrates the preceding creation event which produced the composition concept which was input to *event1_2*. *Event1_3* is the ProgramNotesProduction Event. It needs to be separately defined to ensure that the Agent/Role pair of Brian Large/Note Producer is associated with the "ProgramNotes" output resource. The full version of this paper [22] contains an XML instantiation of this scenario based on the XML Schema in that paper.

6. Next Steps

The modeling concepts described in this paper are the first stage of our work within Harmony to understanding mappings among the metadata schemas from different domains. We close with some observations on how these mappings might work for some representative metadata vocabularies and on the possible mechanisms for performing such mappings.

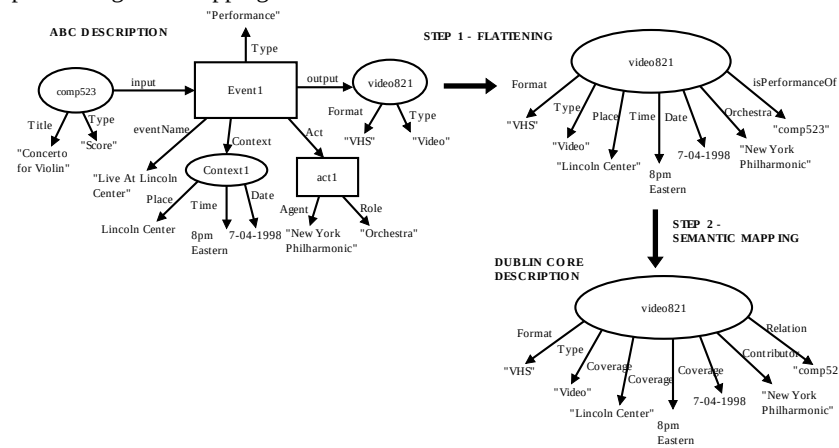


Fig. 6. The 2-step Mapping Process from ABC to Dublin Core

The complexity of the mapping process varies according to the metadata vocabulary. For domains that use a flat unstructured resource-centric metadata model (e.g., Dublin Core, AACR [20]), the mapping process can be broken down into two steps: transformation from the ABC event-aware model to the resource-centric model,

and mapping of the ABC semantic elements to the specific domain's semantic elements. Figure 6 illustrates these steps in mapping from an ABC description of the (simplified) scenario performance to a Dublin Core description of the video. Although the MPEG-7 data model is not explicitly event-aware, it does support the concept of time-based segmentation within audiovisual documents, which reflects the sequence of the original events which were recorded. One approach is to map the ABC model's descriptions of actual real-world events to descriptions of segments within the audiovisual content. Since the CIDOC/CRM and INDECS models both use an event-aware metadata model, it is expected that the structural mapping process from ABC to these schemes (step 1 in Figure 6) will be relatively simple.

There are a number of possible mechanisms available for the mapping process. Some of these are non-procedural, including:

- merging XML Infosets into a single composite Infoset [18, 25];
- using Equivalence classes within XML Schema Language to define mappings [28];
- using XSLT (XSL Transformation Language) [17] to transform an XML description from one domain to another.

We expect, however, that none of these approaches will be able to cope with mapping between the broad range of community-specific semantics which can be “dropped in” within the unifying framework provided by ABC. Recognizing this, we also plan to investigate a number of proposals for a logic language expressed over the RDF data model, which may be useful for this purpose, such as [2, 19].

In the end these investigations and mechanisms will need to take into account a theme common across the metadata field. Expressive power is often desirable for metadata descriptions, but expressiveness comes at the cost of complexity. The success of any model and mechanisms for mapping among multiple descriptive vocabularies will be measured by whether it is feasible to build usable and deployable systems that implement them.

Acknowledgements

The authors wish to acknowledge the contributions to this work by the participants at the ABC Workshop in January, 2000: Tom Baker, Mark Bide, David Bearman, Elliot Christian, Tom Delsey, Arthur Haynes, John Kunze, Clifford Lynch, Eric Miller, Paul Miller, Godfrey Rust, Ralph Swick, and Jennifer Trant. Support for the work in this document came from a number of sources including NSF Grant 9905955, JISC Grant 9906, and DSTC Pty Ltd.

References

- [1] *ABC Workshop*, http://www.ilrt.bris.ac.uk/discovery/harmony/abc_workshop.htm.
- [2] *DARPA Agent Markup Language (DAML)*, <http://www.oasis-open.org/cover/daml.html>.
- [3] *Dublin Core Metadata Initiative*, <http://purl.org/DC>.
- [4] *Dublin Core/MARC/GILS Crosswalk*, <http://lcweb.loc.gov/marc/dccross.html>.

- [5] "Functional Requirements for Bibliographic Records," International Federation of Library Associations and Institutions <http://www.ifla.org/VII/s13/frbr/frbr.pdf>, March 1998.
- [6] *The Harmony Project*, <http://www.ilrt.bris.ac.uk/discovery/harmony/>.
- [7] *INDECS Home Page: Interoperability of Data in E-Commerce Systems*, <http://www.indecs.org/>.
- [8] "MPEG-7 Requirements Document," International Organisation for Standardisation, Requirements IDO/IEC JTC1/SC29/WG11, October 1998.
- [9] S. Abiteboul, P. Buneman, and D. Suciu, *Data on the web: from relations to semistructured data and XML*. San Francisco: Morgan Kaufmann, 2000.
- [10] D. Bearman, G. Rust, S. Weibel, E. Miller, and J. Trant, "A Common Model to Support Interoperable Metadata. Progress report on reconciling metadata requirements from the Dublin Core and INDECS/DOI Communities," *D-Lib Magazine*, 5 (January 1999), <http://www.dlib.org/dlib/january99/bearman/01bearman.html>, 1999.
- [11] D. Bearman and K. Sochats, "Metadata Requirements for Evidence.," Archives & Museum Informatics, University of Pittsburgh, School of Information Science, Pittsburgh, PA <http://www.lis.pitt.edu/~nhprc/BACartic.html>, 1996.
- [12] J. F. Bennett, *Events and their names*. Indianapolis: Hackett Pub. Co., 1988.
- [13] P. V. Biron and A. Malhotra, "XML Schema Part 2: Datatypes," World Wide Consortium, W3C Working Draft WD-xmlschema-2-2000025, <http://www.w3.org/TR/xmlschema-2/>, April 7 2000.
- [14] G. Booch, J. Rumbaugh, and I. Jacobson, *The unified modeling language user guide*. Reading Mass.: Addison-Wesley, 1999.
- [15] D. Brickley and R. V. Guha, "Resource Description Framework (RDF) Schema Specification," World Wide Web Consortium, W3C Candidate Recommendation CR-rdf-schema-20000327, <http://www.w3.org/TR/2000/CR-rdf-schema-20000327/>, March 27 2000.
- [16] D. Brickley, J. Hunter, and C. Lagoze, "ABC: A Logical Model for Metadata Interoperability," Harmony Project, Working Paper http://www.ilrt.bris.ac.uk/discovery/harmony/docs/abc/abc_draft.html, 1999.
- [17] J. Clark, "XSL Transformations (XSLT)," World Wide Web Consortium, W3C Recommendation REC-xslt-19991116, <http://www.w3.org/TR/xslt>, November 16 1999.
- [18] J. Cowan and D. Megginson, "XML Information Set," World Wide Web Consortium, W3C Working Draft WD-xml-infoiset-19991220, <http://www.w3.org/TR/xml-infoiset>, December 20 1999.
- [19] D. Fensel, I. Horrocks, F. Van Marmelen, S. Decker, M. Erdmann, and M. Klein, "OIL in a Nutshell," Vrije Universiteit Amsterdam, Amsterdam <http://www.cs.vu.nl/~dieter/oil/oil.nutshell.pdf>, 1999.
- [20] M. Gorman, *The concise AACR2, 1988 revision*. Chicago: American Library Association, 1989.
- [21] ICOM/CIDOC Documentation Standards Group, *CIDOC Conceptual Reference Model*, <http://www.ville-ge.ch/musinfo/cidoc/oocmodel/>.
- [22] C. Lagoze, J. Hunter, and D. Brickley, "An Event-Aware Model for Metadata Interoperability," Cornell University, Ithaca, Cornell Computer Science Technical Report TR2000-1801, <http://www.ncstrl.org/Dienst/UI/1.0/Display/ncstrl.cornell/TR2000-1801>, June 30 2000.
- [23] C. Lagoze, C. A. Lynch, and R. D. Jr., "The Warwick Framework: A Container Architecture for Aggregating Sets of Metadata," Cornell University Computer Science, Technical Report TR96-1593,

- tr.cs.cornell.edu:80/Dienst/UI/2.0/Describe/ncstrl.cornell/TR96-1593?abstract=, June 1996.
- [24] O. Lassila and R. R. Swick, "Resource Description Framework: (RDF) Model and Syntax Specification," World Wide Web Consortium, W3C Proposed Recommendation PR-rdf-syntax-19990105, <http://www.w3.org/TR/PR-rdf-syntax/>, January 1999.
 - [25] J. Marsh and D. Orchard, "XML Inclusions," World Wide Web Consortium, W3C Working Draft WD-xinclude-20000322, <http://www.w3.org/TR/2000/WD-xinclude-20000322>, March 22 2000.
 - [26] Object Management Group, "OMG Unified Modeling Language Specification Version 1.3," , OMG Specification <http://www.omg.org/cgi-bin/doc?ad/99-06-08.pdf>, June.
 - [27] G. Rust and M. Bide, "The INDECS Metadata Model," <http://www.indecs.org/pdf/model3.pdf>, July 1999 1999.
 - [28] H. S. Thompson, D. Beech, M. Maloney, and N. Mendelsohn, "XML Schema Part 1: Structures," World Wide Web Consortium, W3C Working Draft WD-xmlschema-1-2000225, <http://www.w3.org/TR/xmlschema-1/>, April 7 2000.

QUEST – Querying Specialized Collections on the Web

Martin Heß, Christian Mönch, and Oswald Drobnik

Johann Wolfgang Goethe-University, Frankfurt
Department of Computer Science
60054 Frankfurt, Germany
{hess,moench,drobnik}@tm.informatik.uni-frankfurt.de

Abstract. Ensuring access to specialized web-collections in a fast evolving web environment requires flexible techniques for orientation and querying. The adoption of meta search techniques for web-collections is hindered by the enormous heterogeneity of the resources. In this paper we introduce QUEST — a system for querying specialized collections on the web.

One focus of QUEST is to unify search fields from different collections by relating the search concepts to each other in a concept-taxonomy. To identify the most relevant collections according to a user query, we propose an association-based strategy. Furthermore the Frankfurt Core is introduced — a metadata-scheme for describing web-collections as a whole. Its fields are filled automatically by a metadata-collector component. Finally a prototype of QUEST is presented, demonstrating the integration of the techniques in an overall architecture.

1 Introduction

With the enormous growth of the web and the improvement of web technology one can see that there is a strong trend towards putting content onto the Web in dynamic pages. In fact a large number of documents is stored in web-accessible databases and the content is only emitted in dynamically created pages. It is obvious that databases are more robust, and they allow sites to offer customized content on demand.

The part of the web which is created dynamically has often been referred to as the Invisible or the Hidden Web [16]. The Invisible Web contains specialized, niche information of high quality. Making these resources available in an efficient way is a necessity for future digital libraries.

Most of these databases offer a single search interface on the start page only. Additional fulltext which might characterize the content of the database is rarely provided. Consequently it is very difficult to gather information about these pages automatically. The standard search engines' crawlers such as AltaVista's spider don't find enough information to create comprehensive indexes of the Invisible Web. These high-quality resources are thus ranked inadequately.

Some companies such as Lycos¹, The Big Hub² and Direct Search³ offer catalogues of specialized collections on the web which have been classified manually into a hierarchy of topics.

¹ Lycos http://dir.lycos.com/Reference/Searchable_Databases

² The Big Hub <http://www.thebighub.com>

³ Direct Search <http://gwis2.circ.gwu.edu/~gprice/direct.htm>

A much more convenient way for accessing these collections would be connecting them within a single search interface, applying the common meta search technique. A single search interface is provided to multiple heterogenous back-end search engines. A meta search system sends a user's query to the back-end search engines, combines the results and presents an integrated result-list to the user. Implementing such a meta search engine is hindered by the enormous heterogeneity of the underlying collections. They can differ in media, search fields, quality and size to name just a few criteria.

In the following we sketch some design-issues which have to be considered in future meta search engines for heterogenous collections, as far as the practical integration of large numbers of different collections is concerned.

- *Unifying search concepts*: The different search fields of the collections have to be mapped onto unified search concepts. The mapping should preserve the different semantics of the search fields. For instance the Internet Movie Database (IMDB)⁴ provides two different search fields referring to individuals, one called “People” (persons involved in the creation of a movie) and one called “Character” (name of the individual portrayed in a movie).
- *Selective query-routing*: Broadcasting a query to each underlying collection results in large response times and high demands of net resources. Consequently a query should be routed selectively only to the most appropriate collections.
- *Describing collections*: Identifying the most appropriate collection is strongly dependent on the quality of the information about each collection (metadata). All available information and properties of a collection should be gathered (preferably automatically) and stored in an appropriate metadata-scheme. This scheme has to be powerful enough to combine all characteristics of a collection and simple enough to enhance its acceptance.

Improving access to heterogenous information sources is an active area of research. Academic systems using the mediator approach are Ariadne [2], Tsimmis [5] and Information Manifold [12]. Popular examples of commercial meta search engines are MetaCrawler⁵ and Highway61⁶.

In the next three sections we address these issues, offering practical solutions for each. In Section 5 we prove their feasibility by introducing an architecture of a meta search engine for specialized heterogenous collections.

2 Unifying Search Concepts

Search interfaces of specialized Web-Collections offer individual search options to facilitate access to their documents. In addition to the query-term most collections permit the specification of search concepts to limit the search to a certain concept. Some typical examples for search concepts are author, title, keyword, etc. An important requirement

⁴ IMDB <http://www.imdb.com>

⁵ MetaCrawler <http://www.metacrawler.com>

⁶ Highway61 <http://www.highway61.com>

for the design of meta search engines for specialized web collections is, to preserve the ability to search using these collection-specific individual search concepts.

A meta search engine’s interface needs to handle the syntactic and semantic diversity of all concepts. This can be achieved by mapping individual concepts onto a unified concept scheme. Since the semantics of search concepts differ in their degree of generality, it is obvious that the concepts need to be organized hierarchically. In the simplest case a hierarchical taxonomy can be applied. This model relies on the assumptions that concepts are fully contained in other concepts or contain other concepts fully.

In figure 1 we present an excerpt of our taxonomy of search concepts for specialized collections. Placed on top is the most general concept “thing” from which all less general concepts are derived. The taxonomy is used for identifying subconcepts of a given search concept. If those concepts are supported by the underlying collections, they can be queried by a meta search engine.

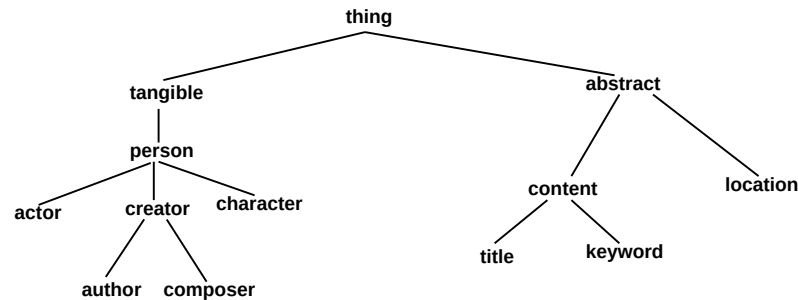


Fig. 1. Concept taxonomy

For example, if a query “person:steven spielberg” is submitted, two queries will be sent to IMDB: “Character:steven spielberg” and “People:steven spielberg” (provided that IMDB’s Character-field is mapped onto the character-concept and the People-field is mapped onto the person-concept).

It is obvious that such simplifications are not always appropriate, for instance consider that a creator does not have to be a person necessarily; it can be an institution or a group of persons as well. Nevertheless our tests have shown that the concept taxonomy is sufficient for the task of querying most web collections. In spite of this we have chosen a tool that allows the specification of more complex relations between concepts in case the need arises. We decided to use description logic, a formal logic whose principal objects are structured terms used to describe individual objects in a domain. It suits our needs and allows the future expansion of the existing concept-specifications. In [19] and [13] applications of description logic are presented in which the retrieval task is supported by specifying domain knowledge. One implementation of a description logic is the CLASSIC-System[3], which allows concepts to be assigned to atomic concepts using Lisp-declarations. Although we evaluate only the inheritance-property of the taxonomy, CLASSIC allows the specification of objects and relations of much higher complexity.

3 Selective Query Routing

Generally meta search engines broadcast queries to all underlying databases. This strategy is not acceptable if specialized collections are concerned. The information stored on each collection is tightly coupled to a specific knowledge domain. Applying the broadcast-strategy would yield a great number of useless requests to collections which might have been excluded preliminarily. Queries have to be routed selectively to the most appropriate collections, according to the information demand the user has expressed. Database selection algorithms and models for databases containing textual documents are described e.g., in [11], [8] and [7]. In [14] a probabilistic framework for database selection is presented.

It is difficult to guess automatically the knowledge domain from a few query terms — especially when considering that most users tend to submit queries consisting only of single terms [9]. This requires the evaluation of domain-specific knowledge provided by thesauri and ontologies. Other strategies apply the statistical analysis of co-occurrence of terms from large numbers of documents to compute association weights indicating thematical correlations ([6],[4]).

In our approach, the gap between the scarce content-descriptions of the collections on the one hand and uncontrolled query-formulation on the other hand, is bridged by computing an association weight between a given query-term and each collection. We obtain those weights by evaluating term frequencies and co-occurrence frequencies generated from a general search engine such as AltaVista. In contrast to co-occurrence frequencies and term frequencies available from secluded document collections, a general search engine is updated frequently and automatically, consisting of a much larger amount of different terms than even the largest thesaurus.

In [17] a query routing system called *Q-Pilot* is presented which performs query expansion to improve the selection of the most relevant specialized search engine. The additional query terms are obtained dynamically by downloading a number of documents from the web and analyzing them for terms which co-occur frequently with the query-term.

3.1 Method

For each collection we store a number of descriptive terms. The terms have been obtained by exploiting publicly available data from the Web, such as existing categorization schemes (Yahoo, Lycos), fulltext which is underlying ingoing hyperlinks and the title of the collection. Stopwords such as internet, database, web, search, etc. are ignored.

For instance the Marx/Engels Search⁷ — a collection containing documents from several popular communists — is described by the terms *marx*, *engels* and *history*.

The association between two terms t_1 and t_2 is computed by the following function:

$$assoc(t_1, t_2) = \left(\log \frac{N}{freq(t_1)}\right) * \frac{freq(t_1, t_2)}{freq(t_1)} * \left(\log \frac{N}{freq(t_2)}\right) * \frac{freq(t_1, t_2)}{freq(t_2)} \quad (1)$$

⁷ Marx/Engels Search <http://search.marxists.org>

N is the total number of web pages indexed by AltaVista (in September 1999: about 193.5 Million documents).

$freq(t_i)$ determines the number of web pages containing the term t_i , or 1 if no page contains t_i .

$freq(t_i, t_j)$ determines the number of web pages containing the term t_i and the term t_j .

Examining the formula $(\log \frac{N}{freq(t_i)}) * \frac{freq(t_i, t_j)}{freq(t_i)}$ in more detail, one can see that the first multiplier is the inverse document frequency of term t_i and the second is the conditional probability that term t_j is co-cited with term t_i , provided that t_i is cited.

For a given query-term q and a collection col which is characterized by a set $D_{col} = \{d_1, d_2, \dots, d_n\}$ of descriptive terms, we compute the overall association-weight $weight(q, D_{col})$ of q with respect to col by summing up the single association weights:

$$weight(q, D_{col}) = \sum_{i=1}^n assoc(q, d_i) \quad (2)$$

3.2 Experiments

To prove the practicability of the method, we performed experiments in which we queried twelve web-collections containing specialized historical documents of different epoches (from ancient Egypt till World War One). Each collection was queried by the same query-terms (mainly names of historical characters or places). For each query the collections were ranked according to the number of results it produced. We further refer to this ranking as the *correct* ranking. The correct ranking was compared to the ranking-list of computed association weights judging the relevance of the collections with each query-term. We will show that the association-based ranking is nearly as meaningful as the ranking by number of results (it should be noted that the association ranking list can be created without querying the collections directly, thus reducing network load).

We examine whether the correct collections are contained in the top d elements of the ranking produced by the $weight(q, D_{col})$ function. n is the number of collections in which a query produces at least one result ($1 \leq n \leq 12$).

In the first diagram of figure 2 queries that have the same value for n are considered in separate graphs (queries with $3 \leq n \leq 5$ and $n \geq 6$ are merged into one graph each).

The best results can be obtained for queries that produce results in only one collection ($n = 1$). 75% of those queries produce association-based ranking-lists in which the single collection is ranked on the top-position ($d = 1$). If the second position ($d = 2$) is considered additionally, the correct collection is found for 90% of the queries. Within the top-three ($d = 3$) the correct collection is identified for all those queries.

For queries that produce results in two different collections ($n = 2$) good results can be obtained as well, but the more results the queries produce, the more the results blur. Nevertheless in most cases, the collection producing the largest number of results is ranked among the first positions. Regarding the second diagram of figure 2 we can see that for 84% of all queries the most important collection (as far as the number of results is concerned) is ranked among the top three.

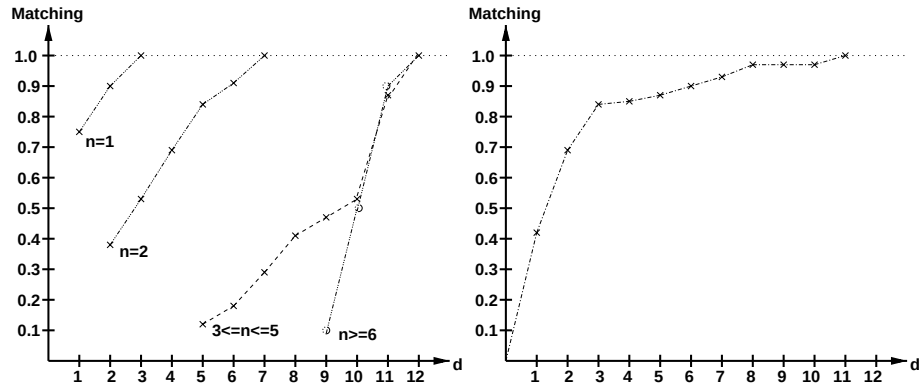


Fig. 2. Matching between association-based ranking and number-of-results ranking

The more specific a query, the better it can be assigned to a special collection. For instance the query “rosa luxemburg” can easily be associated to the Marx/Engels-Search Collection since most occurrences of “rosa luxemburg” in the web refer to the German socialistic feminist. But even less prominent concepts have a presence in the web which is sufficient for evaluation.

Less specific queries such as “alexandria” or “homer” refer to different concepts, producing results in collections that are thematically completely different. Nevertheless analyzing the web for co-occurrences helps to determine the most dominant concept as far as the underlying collections are concerned. For instance the query-term “luxemburg” is still associated strongest with the Marx/Engels-Search Collection, although in the Web the term “luxemburg” refers mostly to the country Luxemburg.

4 Metadata Description

The process of selecting the most appropriate collection strongly depends on the quality of the metadata which is available for each collection. The Dublin Core [18] provides a rudimentary characterization of WWW-resources. But for specialized Web-collections, the 15 Dublin Core elements don’t seem to be sufficient to characterize the collection as a whole.

A metadata description for specialized web collections has to be extensible to integrate additional information available. Next to content-descriptive data this also includes information which allows a qualitative judgement of the collection. Examples of this kind of metadata are the number of accesses in a specified time period, hyperlink reputation⁸, frequency of actualization and the total number of documents accessible by the collection.

Different search engines applying different retrieval strategies can extract and interpret the data available in the metadata descriptions individually.

⁸ This can be judged by simply measuring the number of hyperlinks that point to the resource

Therefore we are developing the Frankfurt Core (FC), a metadata description for specialized collections. The FC is specified using RDF[1]. In addition to the elements of the Dublin Core it contains fields for qualitative judgements:

- *NumberOfDocuments*: Total number of documents accessible by the collection
- *NumberOfIngoingLinks*: Number of hyperlinks pointing to the start-page of the collection
- *NumberOfAccessesPerMonth*: Number of external accesses per month
- *LatestUpdate*: Date of the latest actualization of the collection

and further content-descriptive fields according to the source the descriptive terms have been obtained from:

- *Categories*: List of the most common category-names the collection has been assigned to by existing web-categories (Yahoo, Lycos, etc.)
- *LinkKeywords*: List of the most common terms from the fulltext underlying hyperlinks that point to the start-page of the collection
- *SearchKeywords*: List of the most often queried terms

Another FC-field called *SearchableFields* holds the names of the search fields supported by the collection's search interface and the mapping of those onto the general search concepts provided by the Concept-Taxonomy.

We are also developing a metadata-collector, a tool that collects data available about an individual collection automatically, and that writes it into the appropriate FC-elements. The metadata-collector obtains the information by exploiting existing, publicly available information from the web and by exploiting log-files of the collections (if available). For each field of the FC, an individual strategy for information gathering has to be applied.

The computation of the association weights described in the previous chapter is based on the data from the *Categories*-field and the *Title*-field of the metadata-description. Furthermore the function $weight(q, D_{col})$ can be extended to additionally consider the collections' total number of documents stored in the *NumberOfDocuments*-field (or any other quality metadata mentioned above).

Automatic generation of RDF-metadata is also presented in [10] where a metadata element set, closely related to the Dublin Core, is given to record the characteristics of the resources.

Further strategies to generate characteristic terms of a web-site from the WWW automatically are presented in [15] and [17].

5 Architecture of QUEST

In this section we introduce the architecture of QUEST — a meta search engine for specialized collections on the web which has been realized prototypically at our chair⁹.

⁹ The prototype can be accessed at
<http://www.tm.informatik.uni-frankfurt.de/~cristian/multisearch.html>

The design integrates the specification of unified search concepts from a taxonomy and associative-based selective query routing as described in the previous three chapters.

Each integrated collection is represented by a metadata description containing all information necessary for characterizing and querying the database, such as supported search concepts, descriptive fulltext, contributor of the database and so on. As far as possible, this information has been collected automatically by a Metadata Collector which examines the Web for publicly accessible data about the collections.

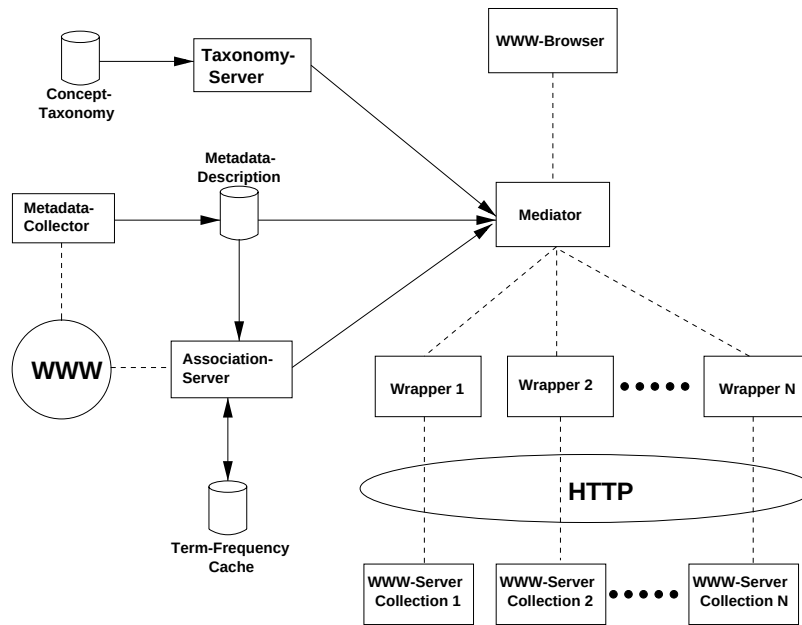


Fig. 3. Architecture of QUEST

The system works as follows: The user submits a query via a Web-Browser. For each term the user can specify a specific search concept from the taxonomy, such as “title:artificial intelligence”, “character:luke skywalker” or “author:marvin minsky”. If no search concept is given the most general concept “thing” is assumed. The query is submitted to the Mediator, whose primary task is to prepare and initiate search requests at the underlying collections.

First the Mediator contacts the Association Server. The Association Server returns a ranked list of association weights, showing the judged relevance of all underlying collections according to the query term. The judging is performed on the basis of term associations between the query term and some collection-specific characteristic terms according to the function $weight(q, D_{col})$ (see section 3.1). The characteristic terms are stored in the metadata description of each collection. To reduce net traffic, determined term-frequency information is cached.

In the next step the Mediator determines which concepts can be queried and prepares to query the collections which have been ranked highest. For each of the selected collections only those search concepts shall be queried which are less general than the one specified, and which are supported by the collection’s search interface.

Therefore the metadata description contains the names of the search concepts supported by the specific collection and their mapping onto the appropriate search concepts of the concept taxonomy. The concept taxonomy is managed by the Taxonomy Server. Given two concepts, the Taxonomy Server can decide if one concept is subsumed by the other.

Each collection is integrated into the meta search engine by its Wrapper. They encapsulate the collection-specific properties. They translate the query into the individual protocol-specific query-string, send it to the underlying collection, and parse the resulting HTML-Document for results. Finally they return a list of results to the Mediator; each result consisting of the URL of the retrieved document and a short snippet.

The distinction between Mediator and Wrapper simplifies the integration of new collections into the system. For each new collection a Wrapper has to be implemented, utilizing the API of the Mediator.

The collection ranking which has been produced by the Association Server can also be used for creating an integrated result list. The higher the collections have been ranked according to the query, the higher the documents retrieved from the specific collection are ranked.

6 Conclusions

In this paper QUEST — a system for querying specialized collections on the web — was presented. One focus of QUEST is to unify search fields from different collections by relating the search concepts to each other in a concept-taxonomy. To identify the most relevant collections according to a user query an association-based strategy is proposed whose feasibility is proved by experiments.

Furthermore the Frankfurt Core is introduced — a metadata-scheme for describing web-collections as a whole. It enhances the Dublin Core with the additional fields containing qualitative and further content-descriptive information. The fields are primarily filled automatically by a metadata-collector-component.

Finally, a prototype of QUEST is presented, demonstrating the integration of the techniques in an overall architecture.

In the future we will extend the existing concept taxonomy for new search concepts and relations, building up a solid domain knowledge of search concepts. This facilitates the integration of further collections and simultaneously allows less restrictive query-formulation. Furthermore we will develop new strategies for automatic metadata generation which can be embedded in the collection description. Improving the process of query routing is another area of research. Therefore we are extending our association-based method to consider boolean queries and further metadata information.

This work was supported by the “Deutsche Forschungsgemeinschaft” (DFG) as part of the project BIB45 FRuv 01–02.

References

1. Resource Description Framework (RDF) Schema Specification, 1999. <http://www.w3.org/TR/PR-rdf-schema/>.
2. J. Ambite, N. Ashish, G. Barish, C. Knoblock, S. Minton, P. Modi, I. Muslea, A. Philpot, and S. Tejada. ARIADNE: A system for constructing mediators for internet sources. In *Proceedings of ACM SIGMOD Conference on Management of Data, Seattle, WA*, 1998.
3. R. Brachmann, A. Borgida, D. McGuinness, and P. Patel-Schneider. 'Reducing' CLASSIC to practice: Knowledge representation theory meets reality. In *KR'92 Principles of Knowledge Representation and Reasoning. Proceedings of the Third International Conference*, pages 247–258.
4. C. Plaunt and B. Norgard. An association-based method for automatic indexing with a controlled vocabulary. In *Journal of the American Society for Information Science, Volume 90 Nr. 10*, 1998.
5. H. Garcia-Molina, Y. Papakonstantinou, D. Quass, A. Rajaraman, Y. Sagiv, J. Ullman, and J. Widom. The tsimmi approach to mediation: Data models and languages. In *Proceedings of Second International Workshop on Next Generation Information Technologies and Systems*, 1995.
6. T. Hofmann. Probabilistic latent semantic indexing. In *ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, USA, 1999.*, 1999.
7. A. L. Powell J. C. French and J. Callan. Effective and efficient automatic database selection. In *Technical Report CS-99-08, Department of Computer Science, University of Virginia*, 1999.
8. Z. Lu J. P. Callan and B. Croft. Searching distributed collections with inference networks. In *Proceedings of the 18th Annual SIGIR Conference*, 1995.
9. B. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: A study of user queries on the web. In *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, volume 32 Number 1, pages 5–17, 1998.
10. C. Jenkins, M. Jackson, P. Burden, and J. Wallis. Automatic RDF Metadata Generation for Resource Discovery. In *Proceedings of the 8th International WWW Conference*, 1999.
11. H. Garcia-Molina L. Gravano and A. Tomasic. The effectiveness of GLOSS for the text database discovery problem. In *Proceedings of the 1994 ACM SIGMOD Conference*, 1994.
12. A. Levy, A. Rajaraman, and J. Ordille. Querying heterogeneous information sources using source descriptions. In *Proceedings of the 22nd VLDB Conference, Mumbai, India*, 1996.
13. R. Möller, V. Haarslev, and B. Neumann. Semantics-based Information Retrieval. In *International Conference on Information Technology and Knowledge Systems, Vienna, Budapest*, 1998.
14. N. Fuhr. Resource Discovery in Distributed Digital Libraries. In *Digital Libraries '99: Advanced Methods and Technologies, Digital Collections, St. Petersburg State University.*, 1999.
15. D. Rafiei and A. O. Mendelzon. What is this Page known for? Computing Web Page Reputations. In *Proceedings of the 9th International WWW Conference*, 2000.
16. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press Books, Addison-Wesley, 1999.
17. A. Sugiura and O. Etzioni. Query Routing for Web Search Engines: Architecture and Experiments. In *Proceedings of the 9th International WWW Conference*, 2000.
18. S. Weibel, J. Kunze, C. Lagoze, and M. Wolf. Dublin Core Metadata for Resource Discovery. <http://www.ietf.org/rfc/rfc2413.txt>, 1998. RFC 2413.
19. C. Welty. DLs for DLs : Description Logics for Digital Libraries. In *Proceedings of the 1998 International Workshop for Description Logics, Trento, Italia*, 1998.

Personal Data in a Large Digital Library

José Manuel Barrueco Cruz¹, Markus J.R. Klink², and Thomas Krichel²

¹ Universitat de València, Universitat de València, Spain

² University of Surrey, Guildford GU2 5XH, England

Abstract. The RePEc Economics library offers the largest distributed source of freely downloadable scientific research reports in the world. RePEc also contains details about Economics institutions, publication outlets and people working in the field. All this data forms a large relational dataset.

In this paper we describe HoPEc, a system that allows to implement access control records personal data within RePEc. The bulk of these records describe the authors of documents. These records are maintained by the authors themselves. We discuss the technical and social aspects of this system.

Paper available online at:

<http://openlib.org/home/krichel/phoenix.a4.pdf>

September 2000

Implementing a Reliable Digital Object Archive^{*}

Brian Cooper, Arturo Crespo, and Hector Garcia-Molina

Department of Computer Science
Stanford University
{cooperb,crespo,hector}@DB.Stanford.EDU

Abstract. An *Archival Repository* reliably stores digital objects for long periods of time (decades or centuries). The archival nature of the system requires new techniques for storing, indexing, and replicating digital objects. In this paper we discuss the specialized indexing needs of a *write-once* archive. We also present a *reliability algorithm* for effectively replicating sets of related objects. We describe a data import utility for archival repositories. Finally, we discuss and evaluate a prototype repository we have built, the Stanford Archival Vault (SAV).

1 Introduction

Information stored and managed by today's digital libraries can be lost within years or decades if special care is not taken. The causes include media and system failures, format obsolescence and bankruptcy of publishers. At Stanford we have implemented a prototype archival repository, the Stanford Archival Vault (SAV, pronounced "save"), for the long term preservation of digital objects. These objects may include documents, their metadata, and the programs for interpreting formats. Our repository does not entirely solve the preservation problem, but we believe it provides an extremely reliable storage infrastructure for preserving digital objects, even as hardware, software, and organizations evolve.

As we implemented and tested our SAV prototype, we identified some unexpected, important challenges that led us to modify our initial design, and to develop some new storage and replication techniques. We believe that the encountered challenges were not unique to our system, but represent some fundamental problems that will be faced in the design of any type of digital library preservation system.

For example, the nature of an archival repository implies that objects should be preserved and not erased. As a result, a repository should not allow users to arbitrarily delete or overwrite digital objects. This *write-once* policy, which is not present in most conventional data stores, forces us to manage data differently. For instance, consider a "set" object that contains pointers to the different materializations of a given document (e.g., the postscript version, the plain text version). The usual way of updating this set is to write a new pointer into

^{*} This material is based upon work supported by the National Science Foundation under Award 9811992.

the set object, or to delete a pointer from the object. Because the write-once policy forbids such changes, managing collections of objects using sets requires new storage structures. Furthermore, these new structures require specifically tailored indexes that can speed up common accesses to digital library sets.

A second area where we faced unexpected challenges was in the configuration of replication “agreements.” Any archival repository must backup its digital objects to remote systems, and hence must enter into some type of agreement with the remote system regarding what objects to replicate. Agreements need to be flexible so that different arrangements can be described, and it is important that new documents be automatically and fully incorporated into the proper agreements, without human intervention. For instance, suppose that a new Physics technical report (TR) is created, consisting of two materializations, a postscript object, and a plain text object. As soon as the “root” digital object for this TR (e.g., the one that links to its components) is added to the set of Physics TRs, all the components should be implicitly added to the proper agreements and automatically backed up. Achieving this flexibility and automation required the concepts of *replication sets* and *annotated links*, concepts that will be useful in any archival repository.

In this paper we discuss the challenges in implementing SAV and the lessons we learned. (A more detailed discussion of some issues can be found in the extended version [5].) We describe the mechanisms that were developed and that could be used in any archival repository. Many of the problems we encountered have been described before in other domains; see Sect. 6 for related work. Here we build on previously developed techniques and, where necessary, present new techniques. Specifically, we make the following contributions:

- We identify the need for an index of the link structure between objects, or *pointer index*. We discuss other important indexes and how, if necessary, these indexes can be built from the pointer index.
- We present a reliability algorithm that replicates digital objects, and detects and corrects corruption in these objects.
- We examine how to use annotated links that restrict traversals over a graph for the purpose of conveniently specifying replication sets. These sets are used by the reliability algorithm, and must grow implicitly for automatic operation of the system.
- We introduce the InfoMonitor, an implemented software package for migrating real-world data (e.g., from a web site) into a repository.
- We present experimental performance results for SAV that illustrate the efficiency and costs of the techniques we describe. Our system scales well to large data sets.

This paper is organized as follows. First, we present a general model for an archival repository in Sect. 2. Then, in Sect. 3 we describe the object storage component of the system. Section 4 discusses the reliability layer, and Sect. 5 presents the InfoMonitor. Section 6 discusses related work.

2 Components of an Archival Repository

Figure 1 shows the architecture of a prototypical archival repository. Our implemented SAV follows this basic design. However, here we address the general principles and features that would form the basis of *any* archival repository. (For specific details of the SAV architecture, refer to [9]).

The architecture in Fig. 1 shows six distinct components of the system. The first component is the *object store*. This component stores and indexes digital objects so that they can be efficiently retrieved by other modules. In addition, the object store manages the assignment of object handles, indexing, and caching of digital objects. The object store can be built on top of an existing storage system, such as a file system or DBMS.

A digital object in our system consists of a list of fields (name/value pairs), and is assigned an object handle. Object handles are used by the system to efficiently locate objects, but are seldom seen by users. Users see human-readable names that are mapped by the system to one or more handles. The name/value pairs are defined by the creator of the object, who generates as many fields as necessary. These fields store content data, metadata, or any other useful values, including another object's handle. Such a *reference field* represents a “link” between two objects that can be used to build complex data structures such as trees, sets, version graphs [32, 17], etc. Other data structures that may be useful are described in [9].

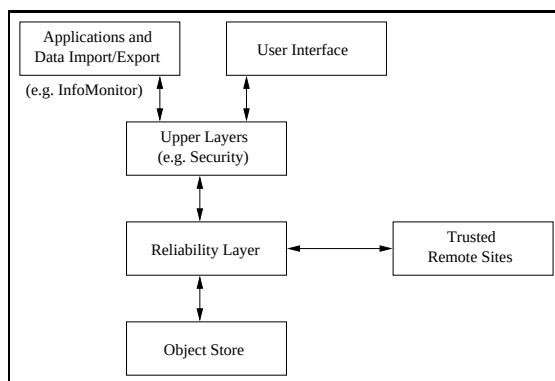


Fig. 1. Architecture of the Archival Repository.

The long term archiving function of the repository is provided by the *reliability layer* in Fig. 1, which manages object replication and corruption detection. This layer relies on different repository sites, usually geographically dispersed, to store copies of the objects. The reliability components at the various sites collaborate in detecting missing or corrupted information, and restoring it. We assume that remote reliability components are trusted. Communications among

trusted reliability components can be encrypted and authenticated using standard techniques. The reliability layer can be configured in various ways (e.g., number of sites involved, number of copies needed for each object) to achieve different levels of reliability and system cost; the determination of appropriate configuration parameters is investigated in [10].

Upper layers on top of the reliability layer provide additional functionality, such as user security, intellectual property management, and query processing. The upper layers provide a programming interface (API) and appropriate information models so that various “applications” can access the repository. One application is an import/export utility that provides batch migration of objects into and out of the repository, from digital libraries that do not provide the high reliability of the archive.

In this paper we focus on the lower system layers (object store and reliability), which are the ones that have been implemented in our initial SAV prototype. However, we do cover one important application: the InfoMonitor (Sect. 5), which migrates information from a standard file system or web site into the repository. In the extended version of this paper [5] we also discuss a user interface for the system administrator that allows him to view the digital objects in the repository, create new objects, group semantically related items together, and construct agreements to replicate objects.

2.1 Archival Repository Properties

An archival repository must enforce certain properties in order to protect digital objects against loss over time. The *write-once* policy specifies that users should not have the capability to delete or modify objects once they are archived. A user can “take an object out of circulation” by changing its access rights, or represent a modification as a new object version, but this is different from physically erasing or changing the object. Allowing users to erase or change objects is dangerous in an archival system, and can introduce ambiguous situations where it is not clear if a object was deleted by a user (and should not be restored) or lost due to some error (and should be restored). Moreover, if objects can be modified, it may not be clear which instance of the object should be restored. Under the write-once policy, the reliability layer simply restores any missing or modified objects, leading to much better long term reliability. Write-once is a policy in SAV, not a requirement of the underlying media as in some other write-once schemes [12] (see Sect. 6 for related work on write-once storage).

The second property is *universal handles*. This property guarantees that an object retains its handle regardless of which repositories it is replicated to, and that the handle is unique within the repository network. Thus, a handle unambiguously identifies a single object. Without this property, the system would have to explicitly record what objects are copies of which, greatly increasing the chances of errors. Moreover, with universal handles, object references can be unambiguously resolved, allowing the structure of a graph of objects to be retained even as the objects are replicated to different sites. The universal handles property also has important efficiency benefits; for example, two sites can

quickly determine whether they have the same objects simply by comparing lists of handles.

3 Object Store

The write-once policy forces us to represent related objects in a way that is unlike traditional data stores. To illustrate, Fig. 2 shows how a “set” can be represented. This set may represent a collection of technical reports, the set of materializations of one report, the set of replication agreements at one site (see Sect. 4), and so on. The set is initially created by generating a “set anchor” A_1 object. An object like O_2 is added to the set by creating a “set member” (represented by M_2 in the figure) which is an intermediate object pointing to both A_1 and O_2 . A member O_2 could be “deleted” (not shown in the figure) by adding a “remove set member object” that links to A_1 and M_2 . All changes are recorded by adding objects rather than by modifying objects.

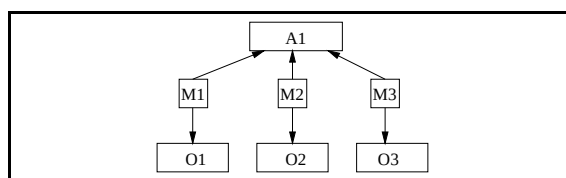


Fig. 2. Structure of set $\{O_1, O_2, O_3\}$

A problem with write-once structures is that they are difficult to traverse. For instance, in order to find all of the members of A_1 , it is necessary to identify the objects that point to A_1 (these objects would be the set member objects, e.g., $M_1, M_2, \dots$, that also point to $O_1, O_2, \dots$). One solution is to scan all repository objects, looking for objects that point to A_1 . Clearly this traversal is very expensive, so we need auxiliary indexes to help us locate objects of interest; these indexes are described in Sect. 3.1. Indexes need to be modified, so they cannot be stored as digital objects, and do not enjoy the high reliability of digital objects. Section 3.2 discusses special mechanisms to ensure the correctness of indexes.

3.1 Indexing Digital Objects

A first critical index is the *handle index* that maps handles to the site-specific identifier (e.g., file name) that locates the object. This index is best implemented as a dictionary (e.g., hash table or balanced binary search tree) with universal handles as keys. This index, like the others we describe, is incrementally maintained. That is, as new objects are created, the index is notified so the appropriate handle-identifier pair is added. The handle index makes universal handles

feasible. Without site-specific information in a handle, and without a handle index, one would be forced to find an object O_1 by scanning all repository objects looking for one with field `HANDLE` = O_1 .

Another important index is the *pointer index* that gives the handles of all objects that link to a given object O_i . For example, for A_1 in Fig. 2, the pointer index can quickly give us the handles for M_1 , M_2 and M_3 , from which we can find the members of set A_1 . Note that in a traditional system a pointer index may be unnecessary if all references are “doubly linked.” However, in an archival repository, A_1 cannot point to M_1 (which was created after A_1). Hence, a pointer index is essential. Again, a pointer index is best implemented as a dictionary. For convenience, the pointer index can be extended to list the outgoing links for each object. This makes it possible to traverse the repository’s graph structure without retrieving the objects themselves.

To make a pointer index feasible, stored fields (Sect. 2) that contain references must be tagged as such. This allows the system to scan repository objects, extract references and build the index. The creator of an object must tag handle fields, either by indicating they are of “handle type” or by using field names that the system recognizes as containing handles (e.g., `PREVIOUS` in our earlier example).

Using the pointer index, a user can efficiently traverse structures such as the set structure in Fig. 2. If necessary for convenience, a specialized “set index” view of the pointer index can be built, which provides operations such as “get set members.” Such a specialized *structure index* could be materialized, or could issue queries to the pointer index to satisfy requests. Similarly, the handle index can be folded into the pointer index, especially if the site-specific identifier (filename) can be computed from the handle itself. For example, if the filename is the hexadecimal representation of the handle, then the list of handles indexed in the pointer index is equivalent to the handle index. Using the pointer index to emulate other indexes does not introduce significant efficiency overhead, but eliminating the pointer index is very expensive for reasons discussed above. Thus, the only index which must be materialized is the pointer index, and other indexes can be materialized to trade space for speed. The implications of this issue in the context of scalability are discussed in Sect. 3.3.

3.2 Maintaining Index Consistency

Indexes are important for the operation of the repository, yet they are inherently not as reliable as digital objects. First, it does not make sense to replicate indexes across sites to achieve reliability. (Indexes contain site specific information that is not useful at the remote sites, and since indexes change often, updating the remote copies would be too expensive). Second, since indexes are updated in place, they are much more prone to software errors than write-once digital objects.

There are two steps to ensure that index errors do not corrupt the underlying digital objects. The first step is to make indexes *disposable*. This means that no information that is critical for the long-term survival of the repository should be placed in an index. In other words, it should be possible to at any time throw

away all indexes and reconstruct correct indexes from the underlying digital objects. As a corollary, all index information must be considered a hint only. For example, if a pointer index tells us that object O_1 points to O_2 , we must verify this (by looking at the actual objects) before performing a critical operation based on this information. An incorrect hint can be inconvenient, especially if the hint is not easily verified, so it is important to keep the indexes as consistent as possible.

There are two ways to ensure the consistency of indexes:

- *Rebuild from scratch*: Periodically discard an index, and completely rebuild it from the objects in the archive. The rebuild procedure is also useful when objects are added in bulk through a data import utility (see Fig. 1).
- *Check and repair*: The information in the index is verified incrementally and, if necessary, corrected.

Note that index rebuilding easily discovers objects that are completely missing from the index, while a check and repair task can only verify existing entries in the index. On the other hand, check and repair allows the index to be available continuously, while the index created by the rebuild task is not available until the rebuild is complete. (Of course, the old, possibly corrupted index could still be used to serve requests while the new index is being built.)

In our implemented SAV system, indexes are kept in main memory and rebuilt from scratch at system startup. They are also rebuilt at the prompting of a user, or at predefined intervals. A check and repair mechanism could be added in the future.

3.3 Performance Measurements

To evaluate the overhead of building indexes, we conducted experiments on our SAV prototype, running on a Gateway E-4200 (450 MHz Pentium III, 256 MB RAM, 128 MB swap, Red Hat Linux 6.0). The SAV is implemented in both in Java 2 and C++; the measurements presented here are from the C++ version. Digital objects containing real documents from the Stanford Database Group's web site were stored in the archive. Five object sets of different sizes were tested in order to assess scalability. The smallest set contained over 54,000 objects and 2 GB of total data, while the largest contained over 270,000 objects and 10 GB of total data. In each set, the average object size was 39 KB. For comparison, the largest data set (10 GB) represents the archived contents of approximately 25 average-sized web sites [7].

The results are shown in Fig. 3. The three lines in the figure represent the three tasks required to rebuild the handle, pointer and set indexes. These tasks are: read objects from disk (solid line), compute the CRCs to detect corruption (dotted line) and index the objects (dashed line). The times scale linearly with the size of the archived data set. The complete index building operation requires an average of 13 milliseconds per archive object (342 seconds per gigabyte), and this time is dominated by the disk read (77%) and CRC computation (21%).

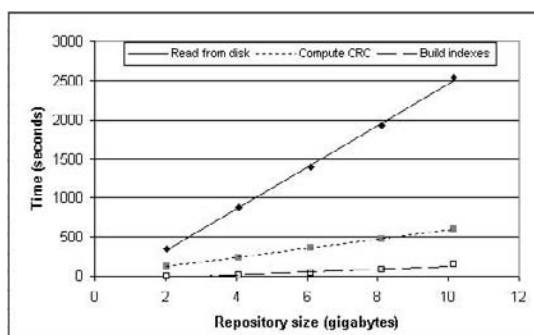


Fig. 3. Performance of the object store.

The high overhead of the disk read and CRC computation is mitigated by the fact that indexes are rarely rebuilt, and most SAV operations avoid these costs by using the indexes after they are already built. Moreover, any scheme that validates indexes by examining the actual objects on disk would incur these costs; our system is not unusual in this respect.

Of course, it is very good that the cost to build indexes scales linearly, but such cost may still be significant for large archives. One solution is to rebuild each type of index at a different time. Another solution is to partition a repository into smaller sets that are reindexed at separate times. This would spread out the rebuilding over time. If this scheme is used, there must be some mechanism to deal with object references that cross partitions, perhaps by querying the indexes for both partitions simultaneously.

It is reasonable to ask how many objects can be indexed before the indexes no longer fit in main memory. We measured the per-object size of indexes as 57 bytes for the handle index, 76 bytes for the pointer index, and 9 bytes for a materialized set index. We assume that, to save space, only the pointer index is materialized (as discussed in Sect. 3.1). If we dedicated 128 MB of RAM to indexes, the SAV could index over 1.7 million objects, or 65 GB of archived data¹. For larger archives, more RAM could be purchased, or the index could be stored on disk and efficiently accessed using known techniques [14].

4 Reliability Layer

As described in Sect. 1, the replication layer backs up objects remotely, detects lost or corrupted objects, and restores them to their pristine state when necessary. The challenge is to develop flexible mechanisms for recording what sites participate in replication agreements, and what objects are backed up where. In addition, we need efficient mechanisms for checking and restoring information.

¹ A large repository may also use compression to save disk space.

In this section we describe the techniques and algorithms that were developed as the SAV prototype was implemented, but that we believe are well suited for any archival repository.

An example will illustrate the basic replication steps we follow. In this example, three sites (e.g., Stanford, Berkeley and MIT) are running SAV instances. The replication process begins when a *replication agreement* is created at Stanford. The replication agreement object identifies the sites that participate (Stanford, MIT, Berkeley) and the objects that are to be replicated. For now, let us assume that this agreement object simply contains pointers to the objects to replicate, a set of technical reports. These technical reports initially exist only at Stanford, so Stanford conducts the first site check. The Stanford site contacts the MIT site and discovers that MIT does not yet know about the agreement, so all of the technical reports and the agreement object are replicated to MIT.² Similarly, all objects are copied to Berkeley.

Each of the three sites then begins a cycle of repeated site checks, connecting to the other two sites and comparing snapshots. As long as there are no errors, the snapshots will agree. However, consider the situation where a technical report is lost at Stanford due to a disk failure. The next site to perform a site check will notice that the report is missing, so a new copy of the report object will be copied back to the Stanford site.

4.1 Constructing Snapshots of the Replication Set

In the above example, we suggested that an agreement object point to the “covered” technical report objects. This is clearly not a good idea since we could never add more objects to the agreement. (The agreement object cannot be modified.) An alternative is to treat the agreement object as a set anchor, so that any object connected via a “set member” object is covered. For example, in Fig. 4, R_2 would cover O_2 and O_3 . (In this figure, please ignore for now the different types of pointers.) This is still not flexible enough, since new objects would have to be explicitly linked to R_2 .

Our solution is to recursively define the covered objects in terms of the link structure of the repository. To illustrate, suppose we wish to cover all versions of a technical report under agreement R_1 in Fig. 4. The different versions of the report, $V_1, V_2 \dots V_n$, are related using a version chain, in which version V_i points to the previous version V_{i-1} . Initially, the first version V_1 is added to R_1 (through M_1). When V_2 is created, it need not be explicitly added to R_1 . Our replication algorithm will implicitly include V_2 in R_1 because there is a path to it from R_1 (via M_1 and V_1). As more versions are created, they are also implicitly included. Thus, the *replication set* of R_1 includes all objects recursively reachable from R_1 (“backwards” links count).

² As described earlier, the reliability layers at each site trust each other, so they willingly take each others’ agreements and objects. Clearly, before the agreement object was created, Stanford checked with the other sites to see if there was enough storage capacity, or to arrange for payment for the service.

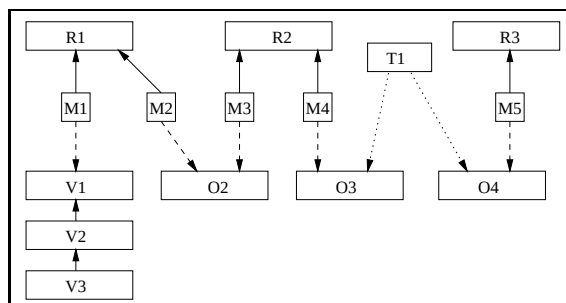


Fig. 4. Example replication sets.

There is a problem with this simple description of a replication set. To illustrate, consider agreements R_1 and R_2 in Fig. 4. Their replication sets are connected by O_2 , so if we blindly include everything that is linked to R_1 in its replication set, we would include all of R_2 's set! Even if agreements do not overlap, other objects may act as *bridges* and connect them. For instance, in Fig. 4, object T_1 is such a bridge object. (Object T_1 may be linking objects written by the same author, for example.)

Our solution is to annotate repository links to indicate when they should be traversed in building replication sets. Some links, like the ones out of T_1 in Fig. 4, should never be traversed. Links such as these have nothing to do with replication, and are shown as dotted lines in the figure. Other links like the ones between M_2 and O_2 , and between M_3 and O_2 , should only be traversed in the direction of their “arrow” to avoid merging replication sets. Such links are shown as dashed lines in the figure. When computing the replication set for R_1 we would reach O_2 but would stop there. Similarly, when computing the R_2 set we would also reach O_2 , but would again stop there.

In summary, we use the concept of a graph with *annotated links*. In such a graph, every link is annotated in one of three ways:

1. *two-way recursive*: The link should always be traversed during a replication set traversal.
2. *one-way recursive*: The links should only be traversed in the direction of the link during a replication set traversal.
3. *non-recursive*: The link should never be traversed when defining a replication set.

The annotated type of a link is specified when the link (and thus the object containing the link) is created. The example shown in Fig. 4 can serve as a template for determining how links should be marked. If it is desirable to change the annotation on a link after it is created, then the replication set traversal algorithm must be extended to allow the annotations on links to be modified by an administrator. Since modifications cannot be written to objects, these modifications can be represented as version chains, and the traversal algorithm

would be designed only to consider the most current version of a link when deciding whether to traverse it. This is an example of the generally applicable strategy of representing modifications as version chains rather than modifying digital objects themselves.

4.2 Detecting Object Corruption

Each site periodically constructs a *snapshot* of the replication set of each known agreement. A snapshot includes the handles of all non-corrupted objects that are part of the agreement. Snapshots are then compared with the corresponding ones at remote sites.

The snapshot construction algorithm is described in detail in [5]. In summary, the algorithm is seeded with the handle of the replication agreement, and traverses the object graph to find other objects in the agreement. The traversal ignores the direction of the object links but is constrained by the link annotations. Any uncorrupted object found during this traversal is considered part of the snapshot.

Once snapshots are created at both sites, the remote site sends the snapshot to the local site, and the local site performs a comparison. Any handles missing locally represent objects that must be retrieved from the remote site, and any handles missing remotely represent objects that must be sent to the remote site.

4.3 Performance Measurements

In order to evaluate the performance of the reliability layer, we conducted experiments on our SAV prototype. We performed the reliability algorithm on the data sets described in Sect. 3.3, using two different SAV instances running on identical machines connected by 10 Mbit Ethernet. The measurements are shown in Fig. 5. In the figure, the solid line represents the time to construct a snapshot

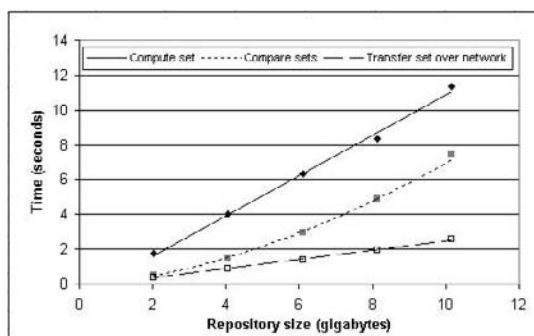


Fig. 5. Performance of the reliability layer.

at a particular repository site. This process must be repeated at both the local and remote sites for each site check; however, the snapshot construction at different sites can run concurrently. The snapshot construction time scales linearly with repository size, and represents an incremental duration of 40 microseconds per object (1.2 seconds per gigabyte).

The snapshot comparison time (dotted line in Fig. 5) increases as $n \log(n)$ (due to our use of a red-black balanced binary tree). This non-linearity is inconsequential since our implementation also examines each object on disk for corruption during the site check, and the I/O cost dominates the time to compare the snapshots by three orders of magnitude. If the examination for corruption is done lazily between (instead of during) site checks, then the comparison time would consume a larger fraction of the site check time (about 1/3 in our experiments). If the non-linear growth of the comparison time hinders performance and scalability, we could substitute a scheme whose time grew linearly, for example by inserting handles into a hash table instead of a binary tree.

The amount of time to send a snapshot from one site to another was 10 microseconds per object (267 milliseconds/gigabyte), as shown by the dashed line in Fig. 5. Various optimizations are possible for use with slow networks or very large repositories, as described in [5, 8].

5 The InfoMonitor

After developing SAV, we discovered a “sad fact” about archival repositories: Many users do not *want* to deposit their digital objects in an archival repository, or in any form of digital library for that matter! They are perfectly happy with their objects residing on conventional file systems or web servers, where they can use their favorite editors and tools to work on them. After all, it is not *their* job to ensure that their objects are available to future generations years from now. However, preservation *is* the job of a librarian, who needs tools to “capture” important objects in a way that does not require active participation by users (but of course requires user consent). The InfoMonitor we describe in this section represents one such tool; the goal is to provide an automated way to migrate data into the archive.

The InfoMonitor serves as a “bridge” between a repository such as SAV and an existing environment where digital objects reside. Our example environment is a web site (but the InfoMonitor can be used in other scenarios too). Users continue to create, edit and access web pages using standard tools (e.g., text editors). The InfoMonitor carefully tracks the files representing the web pages, and decides what objects should be archived. In addition, it monitors changes to the files, *translating* those changes into repository updates.

One of the hardest challenges faced by the InfoMonitor is in deciding how to interpret the changes to the web site. For example, suppose that a web page is modified. Modifications are not allowed on the repository, so the action must be automatically translated into the creation of a new version of the corresponding digital object. If the web page is deleted, a “final” version is added in the repos-

itory, indicating that the web page was removed. Changes to the web site file structure must be carefully analyzed to determine how they impact the archived objects. For instance, if a web page is “moved” from one location to another, this action can be interpreted as a deletion followed by an insertion, or it can be interpreted as new version of the web page (where one of its properties, its file name, was changed).

The InfoMonitor offers an administrative user interface, analogous to the one described in [5]. Through this interface, an administrator can define portions of the web site to archive (by setting “filters”), and can examine archived objects and how they map to web site files. The interface also offers a historical view, where archived objects can be viewed as of a given time. Finally, the administrator can also restore web site files based on the repository objects. Thus, the InfoMonitor offers a fairly automated way to archive a web site. Web users do not need to perform explicit saves to the repository, yet their pages are safely archived.

Figure 6 illustrates how the InfoMonitor represents the web pages as digital objects. The left hand side structure mimics the target file structure, while the right side represents the selection filters and other data. If the top level InfoMonitor Directory is added to a replication agreement, then this entire structure will be replicated at other repository sites.

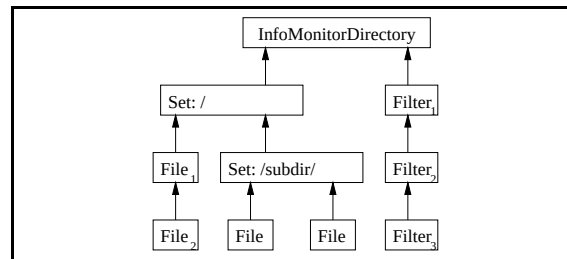


Fig. 6. The InfoMonitor creates this data structure in the SAV.

Initially, the structure of Fig. 6 is created by a bulk load utility that scans the web site. (This same utility was used to acquire the data sets used for the experiments of Sect. 3 and Sect. 4.) The InfoMonitor can perform two types of periodic checks to track the web site: a quick and a slow one. The quick scan compares the timestamps of files with those of the archived objects, to detect new or modified files. Timestamps can be unreliable, so the slow scan actually compares the contents of files to the archived content. In either case, as changes are observed, the appropriate objects are added to the archival repository.

The InfoMonitor has been implemented as part of our SAV prototype. It is currently being used to archive 26,000 files (2 GB) of our group’s web site. Additional details and performance numbers are available in [6].

6 Related Work

The digital library community has begun to focus on the problem of designing and implementing long term archives [15, 11, 13, 30]. Several projects have focused on building archives, including the Computing Research Repository [19], the Archival Intermemory Project [3] and the Victorian Electronic Records Strategy [31]. These projects have focused on different archive architectures than the SAV design we discuss here, and information discovery, not preservation, has been the focus of many of the efforts. The San Diego Supercomputer Center [28] has examined indexing digital archives from the standpoint of metadata; such an infrastructure would be useful as a document discovery mechanism in the “upper layers” mentioned in Sect. 2. The Internet Archive [1] is building a collection of archived web pages, but so far has not addressed the problem of preservation.

The archiving problem is related to the problem of increasing the reliability of file systems. The traditional solution is data backup [4, 21, 23]. Several commercial products use hierarchical replication systems to automatically backup and reliably store data (e.g., [22]). The backup problem focuses on shorter durations than the archiving problem. Moreover, users of backup systems are usually interested in restoring the most current version of data, while archives are responsible for storing all versions. Another approach is to redesign the file system itself to incorporate more reliability features. One idea is to use Redundant Arrays of Inexpensive Disks (RAID) [26], so that disk failures can be overcome. Others have suggested using logs to improve many aspects of the file system, including the reliability [29, 12]. Such systems could serve as the data storage component of our Object Store layer (Sect. 2).

Another related area is the problem of maintaining consistency between nodes in replicated databases. Much work has been done in designing algorithms for propagating data from one replicate to another [2, 27]. These systems focus on systems that allow updates and deletions of objects. Archival Repositories, which do not allow digital objects to be modified or erased, require different approaches. Similarly, filesystems using replication (such as Harp [24], Zebra [20], and others) focus on providing high availability and fault tolerance for frequently accessed filesystems. Our focus is on long term reliability for data that may be archived for decades between accesses.

Finally, many of the issues we discuss here are also present in hypermedia systems [18, 16]. Although the problems are similar, hypermedia systems focus on presentation of objects as much as on the storage of objects, and also must cope with inconsistencies due to modifications and deletions. As a result, the general solutions tend to be similar (e.g., annotating links to restrict graph traversals) although the details and the implementation differ from what we present here.

7 Conclusions

In this paper we have discussed issues that arise when implementing a reliable archival storage system. Although we have discussed these issues from the perspective of our SAV design, they are relevant to the construction of any reliable

archive. We have discussed solutions for defining and indexing digital objects and references between them in a write-once repository. We have discussed efficient algorithms for replicating objects to multiple sites using different replication networks, and for building and comparing snapshots of repository contents so that corruption can be detected. These algorithms allow the set of replicated objects to grow implicitly, rather than through the intervention of a human. We have also described the InfoMonitor, a tool for automatically importing and tracking information outside the repository. The SAV prototype demonstrates that a reliable archive can be built, that it can operate efficiently, and that it can interact effectively with the outside world.

References

- [1] Internet Archive. The Internet Archive: Building an Internet library. <http://www.rlg.org/longterm/index.html>, 2000.
- [2] Yuri Breitbart, Raghavan Komondoor, Rajeev Rastogi, S. Seshadri, and Avi Silberschatz. Update propagation protocols for replicated databases. In *Proceedings of the ACM SIGMOD Conference*, 1999.
- [3] Yuan Chen, Jan Edler, Andrew Goldberg, Allan Gottlieb, Sumeet Sobti, and Peter Yianilos. A prototype implementation of archival intermemory. In *Proceedings of the Fourth ACM DL Conference*, 1999.
- [4] Ann Chervenak, Vivekanand Vellanki, and Zachary Kurmas. Protecting file systems: A survey of backup techniques. In *Proceedings Joint NASA and IEEE Mass Storage Conference*, March 1998.
- [5] Brian Cooper, Arturo Crespo, and Hector Garcia-Molina. Implementing a reliable digital object archive. <http://dbpubs.stanford.edu/pub/2000-27>, 2000. Extended version of paper.
- [6] Brian Cooper and Hector Garcia-Molina. InfoMonitor: Unobtrusively archiving a World Wide Web server. <http://www-db.stanford.edu/pub/papers/fmpaper.ps>, 2000. Technical Report.
- [7] Inktomi Corporation. Web surpasses one billion documents. <http://www.inktomi.com/new/press/billion.html>, 2000.
- [8] Arturo Crespo and Hector Garcia-Molina. Awareness services for digital libraries. In *Lecture Notes in Computer Science*, volume 1324, 1997.
- [9] Arturo Crespo and Hector Garcia-Molina. Archival storage for digital libraries. In *Proceedings of the Third ACM DL Conference*, 1998.
- [10] Arturo Crespo and Hector Garcia-Molina. Modeling archival repositories for digital libraries. In *Proceedings of the Fourth European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, 2000.
- [11] Jean Deken. Writ in water? an exploration of the gap between archival construct and practice in the machine-readable environment. In *Working With Knowledge Conference*, May 1998. Accessible at <http://www.slac.stanford.edu/pubs/slacpubs/7000/slac-pub-7811.html>.
- [12] Ross Finlayson and David Cheriton. Log files: An extended file service exploiting write-once storage. In *Proceedings of the 11th Symposium on Operating Systems Principles*, November 1987.
- [13] National Science Foundation. Workshop on Data Archival and Information Preservation: Executive summary. <http://cecssrv1.cecs.missouri.edu/NSFWorkshop/execsum.html>, 1999.

- [14] Hector Garcia-Molina, Jeff Ullman, and Jennifer Widom. *Database System Implementation*. Prentice Hall, Upper Saddle River, New Jersey, 2000.
- [15] John Garrett and Donald Waters. Preserving digital information: Report of the Task Force on Archiving of Digital Information, May 1996. Accessible at <http://www.rlg.org/ArchTF/>.
- [16] Kaj Gronbaek and Randall Trigg. Design issues for a Dexter-based hypermedia system. *Communications of the ACM*, 37(2):40–49, February 1994.
- [17] Anja Haake and David Hicks. Verse: Towards hypertext versioning styles. In *Hypertext '96*, 1996.
- [18] Frank Halasz and Mayer Schwartz. The Dexter Hypertext Reference Model. *Communications of the ACM*, 37(2):30–39, February 1994.
- [19] Joseph Halpern and Carl Lagoze. The Computing Research Repository: Promoting the rapid dissemination and archiving of computer science research. In *Proceedings of the Fourth ACM DL Conference*, 1999.
- [20] John Hartman and John Ousterhout. The Zebra striped network file system. In *Proceedings 14th Symposium on Operating Systems Principles*, December 1993.
- [21] Norman C. Hutchinson, Stephen Manley, Mike Federwisch, Guy Harris, Dave Hitz, Steven Kleiman, and Sean O'Malley. Logical vs. physical file system backup. In *Proceedings of the Third USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 1999.
- [22] Tivoli Systems Inc. Tivoli storage manager. http://www.tivoli.com/products/-index/storage_mgr/, 1999.
- [23] Richard P. King, Nagui Halim, Hector Garcia-Molina, and Christos A. Polyzois. Management of a remote backup copy for disaster recovery. *TODS*, 16(2):338–68, 1991.
- [24] Barbara Liskov, Sanjay Ghemawat, Robert Gruber, Paul Johnson, Liuba Shrira, and Michael Williams. Replication in the Harp file system. In *Proceedings 13th Symposium on Operating Systems Principles*, October 1991.
- [25] Stanford Conservation Online. Electronic storage media. <http://palimpsest.stanford.edu/bytopic/electronic-records/electronic-storage-media/>, 2000.
- [26] David Patterson, Garth Gibson, and Randy H. Katz. A case for redundant arrays of inexpensive disks (RAID). *SIGMOD Record*, 17(3):109–116, September 1988.
- [27] Michael Rabinovich, Narain Gehani, and Alex Kononov. Efficient update propagation in epidemic replicated databases. In *Proceedings of the 5th International Conference on Extending Database Technology*, 1996.
- [28] Arcot Rajasekar, Richard Marciano, and Reagan Moore. Collection-based persistent archives. <http://www.sdsc.edu/NARA/Publications/OTHER/Persistent/-Persistent.html>, 2000.
- [29] Mendel Rosenblum and John K. Ousterhout. The design and implementation of a log-structured file system. In *Proceedings 13th Symposium on Operating Systems Principles*, October 1991.
- [30] David Rosenthal and Vicky Reich. Permanent web publishing. <http://lockss.stanford.edu/>, 2000. To appear at Freenix, San Diego, CA, June 2000.
- [31] Victorian Electronic Records Strategy. Victorian electronic records strategy final report. <http://home.vicnet.net.au/~provic/vers/final.htm>, 1999.
- [32] Walter Tichy. RCS - a system for version control. *Software - Practice and Experience*, 15(7):637–654, 1985.

Policy-Carrying, Policy-Enforcing Digital Objects

Sandra Payette and Carl Lagoze

Department of Computer Science, Cornell University,
Ithaca, NY, USA
{payette, lagoze}@cs.cornell.edu

Abstract. We describe the motivation for moving policy enforcement for access control down to the digital object level. The reasons for this include handling of item-specific behaviors, adapting to evolution of digital objects, and permitting objects to move among repositories and portable devices. We then describe our experiments that integrate the Fedora architecture for digital objects and repositories and the PoET implementation of security automata to effect such object-centric policy enforcement.

1 From Traditional Access Control to Flexible Policy Enforcement

Access control in digital libraries must be scalable, flexible, and extensible – accommodating a wide range of objects and usage scenarios. This stands in contrast to traditional access control models, which are limited to a relatively fixed set of operating system abstractions (e.g., files) and computing actions (e.g., read, write). The limitation of these models has motivated the programming language, operating systems, and security communities to develop alternative and richer methods, including an approach known as language-based security [1].

The shortcomings of traditional access control schemes are particularly acute in digital libraries where idiosyncratic and complex digital objects are the norm rather than the exception. Such objects may aggregate mixed content (e.g., text, images, audio), may be highly interactive (e.g., a distance learning module), or may be portable (e.g., an object that is disengaged from its “home ship” for use on a mobile device). A typical security model has been to allow/disallow access to particular objects by particular subjects using system-level enforcement mechanisms. These mechanisms do not easily adapt to handle the wide range of scenarios pertinent to particular applications and complex objects. Designed to express only a limited set of access control rules, such mechanisms are often constrained in their *policy enforcement* capabilities.

In this work, we recognize a continuum of policy enforcement – ranging from repository or system-centric to object-centric. This model accommodates general system-wide requirements, such as those covered by standard access control mechanisms, along with more refined requirements that pertain to various types of content objects.

This spectrum can be understood through the metaphor of a museum attempting to enforce a variety of policies to protect its assets. In this hypothetical museum, a security guard is employed to stand guard at the main entrance and make decisions about who can enter for what reasons. The guard enforces some general rules of entry (e.g., people must have paid an entrance fee, people must check their bags in the cloak room), as well as policies that ensure the safety of the building (e.g., anyone carrying a weapon is denied entry). The guard may permit access to private rooms (e.g., by providing a special pass), but once a person enters the room, the guard cannot control what the person does. In terms of our continuum, the guard is a mechanism for enforcing museum-wide rules, typical of the system-centric end of our spectrum.

Carrying our metaphor forward, what happens once people have passed through the main entrance? We wouldn't expect the guard to have the knowledge or ability to oversee specific activities that occur inside the many rooms of the museum. Each room may have its own protocols, rules, and restrictions for specific behaviors. Also, room-specific rules may evolve on a day-to-day basis, or they may address the occurrence of special events. To be effective, the main guard would have to be omniscient and omnipresent, and equipped with all of the rules and tools for every unique situation.

As a better alternative, the guard should delegate control to others who enforce room-specific policies locally. A special exhibition manager would be equipped to enforce policies unique to a particular exhibition room. For example, she would monitor activity, allowing (1) members of the general public to touch certain items if they wear a white glove, (2) scholars with proper credentials to gain entrance to a private alcove to see selected rare documents, and (3) curators to open display cases to insert or remove items. Such policy enforcement is both object and context-specific.

Access control in digital libraries presents a challenge similar to that of the museum. Digital libraries contain a rich array of digital objects for which they must provide secure enforcement of both general-purpose and context-specific policies. This paper describes our work at Cornell to provide such secure policy enforcement.

2 Research Context

Our investigation of policy enforcement for digital libraries builds on and extends two ongoing research programs at Cornell: (1) distributed architectures for digital objects and repositories, and (2) enforceable security policies.

In the Fedora project, we have demonstrated interoperability and extensibility for digital objects and repositories [2,3]. The main result of this work is a flexible digital object model – one that supports the aggregation of mixed digital content and the association of rich and customized behaviors with those aggregations. For example, book objects have behavioral access points such as “Get Table of Contents” and “Get

Page.” Multimedia learning objects have access points such as “Get Lesson Video.” While Fedora enables the creation of such customized access points on objects, it still provides for interoperability among varying objects. Through a set of *generic* behaviors common to every digital object, Fedora enables clients to uniformly discover and invoke the customized, content-specific behaviors associated with the object.

Having demonstrated content-specific and extensible behaviors for digital objects, our next goal is to devise an equally flexible access control scheme to support them. The requirements for such a solution are:

- Flexible policy definition: ability to define policies that are either general-purpose or specific to particular digital object behaviors
- Highly secure enforcement: policy enforcement mechanism must be based on well-grounded security techniques
- Conducive to mobile code and mobile computing environment: policies must be enforced on any code that is obtained from the network. Furthermore, policies and enforcement mechanisms must permit digital objects to move among repositories and to mobile devices.

This paper describes an approach for fulfilling these requirements by applying and extending new work in language-based security. Specifically, we are investigating the use of *security automata* for specifying access control policies that are flexible, fine-grained, and *enforceable* [4]. This special class of automata can be used to model policies in the form of finite-state machines. For example, using a state transition diagram, Figure 1 depicts an automaton for the policy “after viewing descriptive metadata, only Cornellians can access the Lesson 1 video (of a distance learning module).”

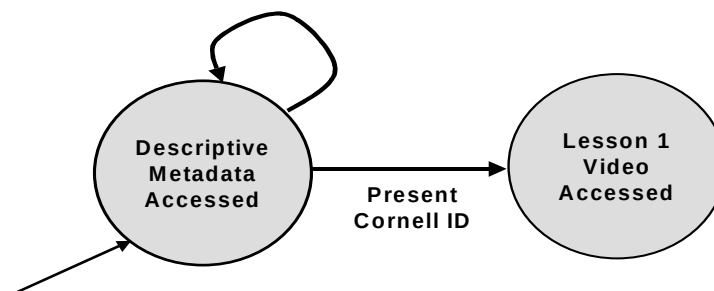


Fig. 1. Simple Security Automata for a Lecture Object Policy

Security policies modeled in this manner can be enforced by mechanisms that simulate an automaton. Essentially the automaton will monitor the executions that occur in a

target system (including application-level and object-level executions). It will prevent the running of any executions that violate a system policy. This approach is demonstrated in Cornell’s Policy Enforcement Toolkit (PoET). In PoET, security automata simulations are implemented as in-line reference monitors (IRM) for the Java environment [5,6]. Through our experiments with PoET and Fedora, we have demonstrated the power of IRM in providing flexible and secure access control for behaviorally-rich digital objects and repositories. The details of the Fedora and PoET technologies will be described in Section 4.

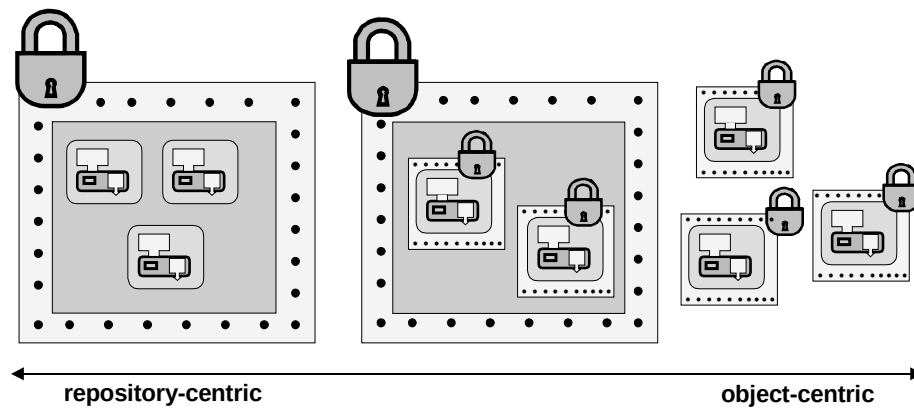


Fig. 2. Policy Enforcement Continuum

As illustrated in Figure 2, our integration of the Fedora and PoET technologies can provide policy enforcement for a range of digital library scenarios. We depict a policy enforcement continuum (from repository-centric to object-centric) where the dotted areas with “locks” indicate the access control target points. On the left, a Fedora repository enforces general-purpose policies over all its contained objects. In the middle, a hybrid approach is used, where some policy enforcement occurs at the repository, while some policies apply to individual digital objects. At the right end of the continuum are digital objects that handle their own policy enforcement. The experiments we report on in this paper are those focused toward the object-centric end of the spectrum. To demonstrate this object-centric approach to policy enforcement, we have designed the *Policy-Carrying, Policy-Enforcing Digital Object*. The next section discusses the motivations for pushing policies and enforcement mechanisms into digital objects.

3 Policy-Carrying, Policy-Enforcing (PCPE) Digital Objects

Most digital libraries will employ a combination of repository-level and object-level policies. The following requirements indicate that the object-level end of the spectrum deserves special attention.

3.1 Behavior-Oriented Policies

Policy definition should be able to address the characteristics, behaviors, and usage scenarios of actual digital library objects - not just system abstractions such as files, directories, and operating system commands. Our Fedora work demonstrates a model whereby behaviors of digital objects correspond to behaviors “real-world” entities. These behaviors are implemented as methods that essentially hide the complexities of underlying data, structure, and file formats that are encapsulated in a digital object container. Referring to our book example, methods such as “Get Table of Contents”, and “Get Page” provide access to content, whether a book is stored as a postscript file, or as a set of digital images.

In Fedora, clients invoke such methods to obtain different “views” of digital objects. To control access to digital object content, we must be able to write policies that correspond to these defined views. In other words, the semantics of policies should parallel the semantics of digital object behaviors. This allows policy definition to be closely aligned with people’s understanding of their content, not of a computer system.

3.2 Specificity vs. Generalization

Although we may be able to identify certain repeatable “design patterns” for policies in digital libraries, we should recognize that particular combinations of these types of policies are often unique to particular objects. This corresponds to the fact that digital resources are generic in some ways, but specialized in other ways.

There are circumstances in which policies should be applied to digital objects in a generic manner. Consider a repository that has a variety of digitized books and lectures pertaining to computer science. There may be a set of *general-purpose* policies pertaining to the collection as a whole. An example of such a policy is: “Cornellians with repository manager status can add and delete objects in the collection.” This policy is quite coarse in nature, and applies equally to all objects in the repository regardless of their individual characteristics. These policies are typically enforced by a repository-wide mechanism.

However, repository-level policy enforcement becomes unworkable for policies that address the idiosyncratic nature of individual digital objects. General-purpose enforcement mechanisms may not be expressive enough to support the variety of variables, conditions, and desired actions pertaining to particular content items. As an example of such an item-specific policy, consider an object that provides access to a digitized lecture. Table 1 outlines a plausible policy governing access to such an object.

Table 1: Policy for Lecture Object “A”

Policy Rules:	
•	Only users who present Cornell credentials, or who pay a \$5 fee, can access the HIGH-RESOLUTION VIDEO.
•	All Users can access SLIDES 1-10 without restriction.
•	Only users who present Cornell credentials can access SLIDES 11-20.
•	All users can access DESCRIPTIVE METADATA without restriction.

Enforcement of such a policy requires a mechanism that can detect requests that correspond to the access points specific to the lecture object, and ensure that the conditions of the policy are not violated. Although one could attempt to extend repository-level mechanisms to deal with these kinds of policies, it is probable that the mechanism would eventually become awkward and overburdened as it tried to accommodate increasing numbers of customized policies. Such object-specific policies are better dealt with in a modular fashion, and linked to the object and its individual access points, instead of a repository.

3.3 Decentralized Policy Management

We envision digital libraries with distributed responsibility for policy definition. In this scenario, system managers implement and maintain policies to ensure security for entire repositories. However, other information stewards have responsibility for defining and maintaining usage policies for sets of digital objects. As we move policy definition to the object level, we can delegate responsibility for policy management to agents who are very familiar with certain classes of objects, or who have a vested interest in managing certain collections.

A critical component of such distributed policy definition is a high-level policy language. There are a number of ongoing projects that are developing such languages, including Cornell’s PoET, the XrML initiative [7], and PolicyMaker [8]. From a digital library perspective, such languages must be intuitive and easy, so that authors and librarians can write policies tailored to particular digital objects, or groups of objects.

3.4 Extensibility

If we take extensibility as a key requirement for digital objects, we are presented with the challenge of evolving policies with the evolution of an object. One of the advantages of the Fedora architecture is the ability to enhance existing objects, over time, by endowing them with new behaviors. Such behavior evolution is facilitated by the modular design of Fedora Digital Objects, and particularly by the use of “pluggable” mechanisms that can add or transform object behaviors. The principle of modularity should be applied to policies and enforcement mechanisms as well. When object behaviors or interfaces change, policies and enforcement mechanisms must easily adapt. Attempts to define and manage monolithic policies at the repository-wide level

would likely be difficult, especially since we cannot anticipate ahead of time the manner in which policies may need to be changed to support new object-level behaviors.

3.5 Portability and Mobile Computing

Portability of digital objects is important for a number of reasons. First, fast and reliable access to objects at a global level often requires caching or mirroring. Second, as we move from a world of fixed networks to one of mobile devices with intermittent connectivity (nomadic devices), we will increasingly confront issues of access to digital content that is copied out of the networked infrastructure. In both cases, we would like digital objects to be secure in transit and to facilitate their own policy enforcement at their target destinations.

4 Experiments: PCPE Digital Objects Using Fedora and PoET

Given the motivations just described, we devised set of experiments to demonstrate Policy Carrying, Policy-Enforcing (PCPE) Digital Objects. In preparation for the experiments, we evolved the Fedora model to support an augmented Digital Object - one that carries its own policies within. As mentioned earlier, the security automata approach provides significant flexibility for policy definition and enforcement, making it an appropriate technology for addressing the unique requirements of digital libraries. Thus, by extending Fedora, and applying Cornell's PoET software, we have stretched the capabilities of our digital object architecture, and have simultaneously tested the flexibility of PoET against real-world requirements of the digital library community. A brief review of the both Fedora and PoET will help illuminate the PCPE experiments.

4.1 Technical Overview: Fedora

Fedora is a modular architecture built on the principle that interoperability and extensibility is best achieved by the clean separation of data, interfaces, and mechanisms [2,3]. A Fedora Repository provides a general-purpose management layer for *Digital Objects*. In their simplest form, Digital Objects are containers that aggregate mime-typed streams of data (e.g., digital images, XML files, metadata), known as *DataStreams*. It should be noted that DataStreams can be references to external data - either disseminations of other Fedora Digital Objects, or service requests to remote data sources. These are known as *ReferenceDataStreams*, and are one way that Fedora enables the aggregation of distributed content. Clients can interact with Fedora Digital Objects through a generic set of methods, known as the *Primitive Disseminator*. These methods can be thought of as the default API for Digital Objects, providing a well-defined open interface for all objects.

In addition to behaving in a generic manner, Digital Objects must be able to mirror real-world entities by providing access methods that make an object behave in a content-specific manner. For example, a natural behavior for a book would be “Get Table of Contents.” Fedora allows the association of rich and extensible behaviors with Digital Objects by “plugging in” generic components known as *Disseminators*. Each Disseminator will aggregate references to: (1) a formally defined behavior interface that defines a set of methods for a particular kind of digital library resource (e.g. Book interface, Lecture Browser interface), (2) an executable mechanism (i.e., a Servlet) that runs these methods, and (3) DataStreams the Servlet should use to fulfill specific method requests. These interfaces and mechanisms are themselves disseminated by Digital Objects, laying the foundation for unlimited extensibility of the architecture. A major strength of the Fedora extensibility model is that clients can use the generic methods (of the default API) to discover and invoke content-specific methods defined on Disseminators. The Digital Object facilitates the invocation of these extended methods, returning customized disseminations of content to the client.

Figure 3 depicts a Digital Object that provides access to a multimedia lecture. There are multiple DataStreams in the object’s core: a low-resolution video of a lecture. (Video-L); a high-resolution video (Video-H); a set of accompanying slides (Slide-1, Slide-2); and an XML file that encodes both synchronization and descriptive metadata. Two Disseminators define the access methods for this object. One defines an interface for experiencing the lecture, the other an interface for obtaining Dublin Core metadata. Although it is not apparent in the diagram, each of these Disseminators references a Servlet - a piece of java code that may be local to the repository, or may be a piece of mobile code. Clients interact with the Digital Object through its generic API methods.

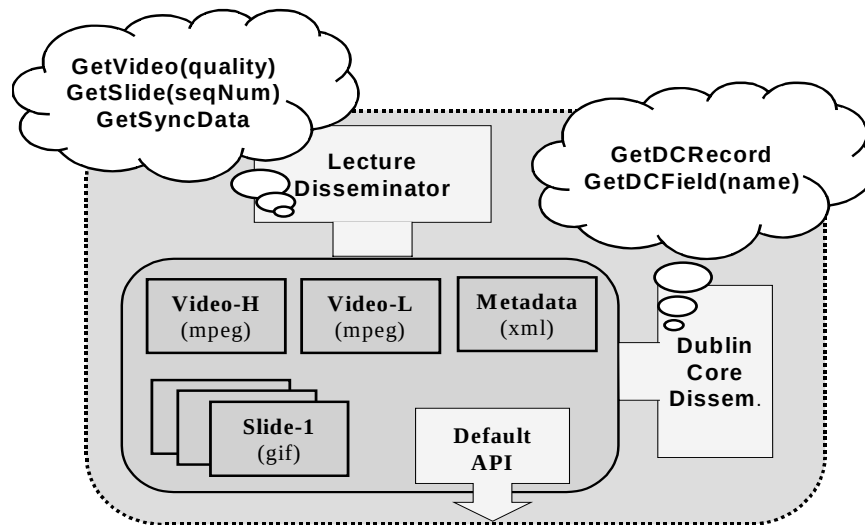


Fig. 3. A multimedia lecture represented as a Fedora Digital Object

Providing access control for these complex objects, without compromising flexibility and extensibility, presents a significant challenge. Early Fedora designs specified Digital Objects with their own internal access control mechanisms, known as Access Managers. The Access Manager was envisioned as a watchdog, overseeing the execution of all methods that could be invoked on a particular Digital Object. The power and flexibility of the PoET software has made it possible for us implement the Access Manager abstraction using the sound underpinnings of security automata theory. An overview of the PoET technology lays the groundwork for the rest of this section.

4.2 Technical Overview: Policy Enforcement Toolkit (PoET)

PoET is an implementation of in-lined reference monitors (IRMs) for Java applications. To review, IRMs can simulate security automata; as such, they can enforce a rich class of access control policies. In general, a reference monitor enforces security policies by mediating all executions that pertain to the policy it is enforcing. A reference monitor must be transparent to applications, and must be protected from attack by applications. IRMs achieve this through bytecode modification by a trusted program rewriter.

In PoET, java applications are converted to *secured* applications by a code rewriter that embeds checks into Java Machine Language programs (bytecode). An easy way to understand this is to imagine PoET "baking" the bytecode in such a manner that the policy becomes ingrained within it. The embedded checks terminate forbidden executions *before* they violate the security policy for the application. To ensure the integrity of the checks, and to increase performance of the modified code, PoET performs program analysis and optimization during the rewriting process. It should be noted that since PoET modifies bytecode - *not* Java source code - the Java compiler is *not* part of the trusted computing base. It follows, then, that a computer running PoET does not require the presence of a Java compiler, nor Java source code for the application.

In PoET, security policies are specified in the *Policy Specification Language* (PSLang), which is a simple, event-oriented language that is somewhat similar to Java in syntax. Policies that are modeled as security automata can be easily expressed in PSLang.

Figure 4 shows the process by which PoET creates a secure application. First, PoET intakes both the policy to be enforced, and the code to be protected. The PoET rewriter decomposes bytecode (Java class files), embeds policy-checking code, performs integrity checks, and generates a modified versions of the class files. All original functionality is preserved, but the resultant class files have the policy checks merged into them. Finally, the PoET class loader transfers the modified code into the Java Virtual Machine (JVM) where it is run as a secured application.

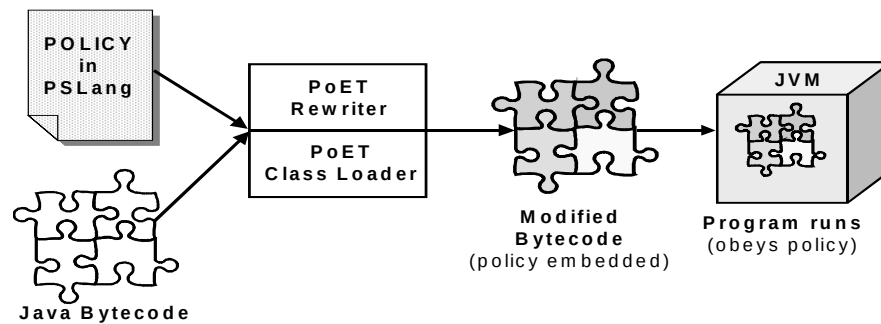


Fig. 4. PoET embeds a policy into Java bytecode

4.3 Integration of Fedora and PoET

Because the Fedora implementation is written as a Java application we were able to take full advantage PoET's in-line reference monitoring. The standard model for PoET is to statically embed policies into application code. Since Digital Object behaviors may evolve, the policies that correspond to them must also evolve. If policies are modified or extended, these changes must be immediately reflected in the executions of Fedora Digital Objects. To ensure this, we enhanced PoET so that it now accepts and modifies pieces of Java bytecode on a "just-in-time" basis. This was particularly important because Digital Objects often depend on shared pieces of mobile code to execute behaviors.

Fedora now associates a *default policy* with each Digital Object in a Repository. The default policy is a *shared* policy that any Digital Object can point to. The default policy specifies access restrictions for the generic API methods common to all Digital Objects. Typically, repository managers will define a default policy that prevents end-users from modifying the contents of Digital Objects, but allows users to browse and obtain disseminations from them.

Additionally, Fedora now supports the association of one content-specific policy with each behavior interface (e.g., a book interface that includes a number of methods). These policies can be customized to a particular Digital Object, or they can be "group policies" that any Digital Objects can associate with. Customized policies are stored as DataStreams inside Digital Objects. Group policies reside in the repository and are stored *by reference* in Digital Objects (i.e., stored as ReferenceDataStreams).

To provide policy enforcement using in-line reference monitoring, Fedora interfaces with PoET. In Fedora, extensible code modules (i.e., Servlets) are used to implement Digital Object behaviors. Originally, Fedora's *Mechanism Manager* obtained these code modules – either locally or from the network – and would load them into the JVM as needed. As part of our integration, we adapted the Mechanism Manager to

interface directly with PoET. Now, whenever a Digital Object receives a request, the Mechanism Manager obtains the requisite code module and sends the appropriate policy and code to the PoET rewriter. From there, PoET takes over to ensure the secure modification of the Java bytecode, and secure loading of the modified classes into the JVM.

4.4 Experiment Results

In collaboration with our library colleagues, we are designing policy enforcement experiments that address policy requirements for a range of object types and institutional contexts. In this paper, we report on our first set of experiments with Cornell's Computer Science lecture archive. Figure 5 depicts a lecture packaged as a Fedora Digital Object secured with PoET's in-line reference monitoring. It should be noted that the diagram does not convey the fact that repository-wide policies are normally in place to prevent untrusted programs from compromising the host computing environment. In the diagram, the Lecture Digital Object introduced earlier now contains two policies: (1) a policy pertaining to the Digital Object's default API behaviors (labeled "Default Policy"), and (2) a customized policy pertaining to the behaviors of the Lecture Disseminator (labeled "Policy-L"). The default policy restricts access to operations that can modify the contents of the object. The Lecture Disseminator policy (Policy-L) is a PSLang specification of the policy we saw earlier in Table 1. Although other lecture objects are structurally similar, they contain different content and different policies.

In support of the object-centric view we have proposed herein, we have achieved three main results, namely the ability to: (1) write highly customized policies tailored to the specific nature of object content, (2) enforce radically different policies for different objects of the same type, and (3) dynamically apply policies to code modules (mobile java programs) that are shared by multiple objects, demonstrating that the same application code can adapt to the policies pertaining to a particular run-time context.

5 Discussion

Object-centric policy management may have its drawbacks. Is it too hard to associate policies with every object? It depends on the nature of the objects, how unique they are, and how uniform policies tend to be. However, as noted earlier, repository-centric policy management has serious scalability and extensibility issues. Our future experiments will focus on the trade-offs of each.

In addition, there are significant mobile computing challenges yet to be addressed with PCPE digital objects. What if a digital object is taken out of its home repository and downloaded onto a mobile computing device? Such objects have moved outside our trusted computing base. For such *portable* objects there is a level of assumption

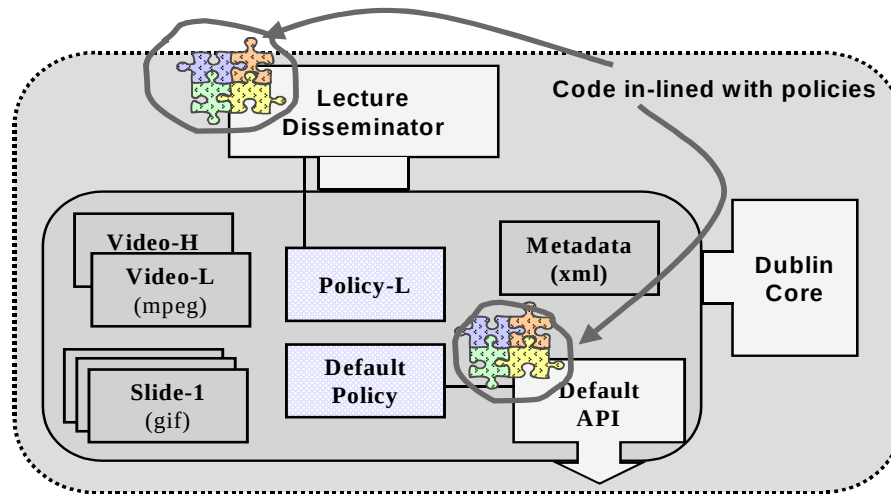


Fig. 5. "Lecture A" Digital Object with Restrictive Policies

around the "receiving" environment. First, the receiving environment must be able to interpret and "run" the digital object. In our experiments, we assume the presence of a Java Virtual Machine and software capable of interpreting Fedora objects. Even if policies are embedded within bytecode, an object may still be vulnerable. A portable digital object must be tamper-proof. In our Fedora/PoET implementation, a digital object that is disembodied from its home repository could be compromised by a savvy attacker who knows how to decompose and modify java bytecode in a manner that allows a policy to be bypassed. Although this is a difficult maneuver, we believe that our PCPE digital objects should be supported by supplemental security measures - possibly the use of public key infrastructure or other trust schemes to support digital object mobility. The digital objects should be secured containers, similar to that approach taken by IBM Cryptolopes [9]. Trust and security for mobile computing are active research areas that will be a continuing thread of inquiry in our digital library research.

5.1 Future Work

Under the auspices of Project Prism [10], we will continue to test increasingly complex policy scenarios in collaboration with Cornell University Library. One area of focus will be in contention between repository policies and digital object policies - how does the overlay or interaction of policies get resolved? A particularly interesting line of future research will be in the area of "intentional" policies. At this time, we have demonstrated the pre-defined binding of particular policies to particular digital objects. In our future work we will study dynamic binding of policies to "virtual subsets" of digital objects. We look to dynamic rule-based policy binding where criteria

can be established that set up the conditions in which objects should be subject to particular policies. This would allow us to say at some arbitrary point in time that "all objects authored by Bill Arms are now subject to the provisions of Policy XYZ."

5.2 Related Work

The Smart Object Dumb Archive (SODA) architecture promotes a similar notion to our PCPE Digital Objects by pushing functionality into objects [11]. Although SODA objects may contain terms and conditions data, it is not clear whether these are expressed in a formal policy language, or what security techniques are used to enforce them. Also, the SODA work does not specifically address issues of flexibility in policy expression and adaptability of enforcement mechanisms.

Others have developed policy specification languages that can be used in conjunction with trust management systems. Keynote [12] has its own policy language that is used in conjunction with the Keynote interpreter. XrML [7] is an XML-based markup language for describing rights, fees, and terms for digital content. XrML does not prescribe an enforcement mechanism, but a way to specify rights information so that various systems can read and interpret it. Cornell's PoET is an application of language-based security. There is no central trust server that grants or denies access. PoET's policy language (PSLang) is optimized for use by a program rewriter that actually modifies application code (java bytecode) to ensure that policies are enforced. The modified programs will prevent executions that violate policy. Relative to other approaches, this technique is particularly useful in securing mobile code applications.

Acknowledgements:

We acknowledge the contributions of Fred Schneider for security automata theory and PoET software; Ulfar Erlingsson for PoET software; Naomi Dushay for work on the Fedora reference implementation; Sugata Mukhopadhyay and Jill Newman for the lecture object testbed. Work described here was supported by NSF Grant IIS-9817416 and ARPA/RADC grant F30602-96-1-0317, AFOSR Grant F49620-94-1-0198, DARPA and Air Force Research Laboratory, Air Force Material Command, USAF, under agreement number F30602-99-1-0533, NSF Grant 9703470, and a grant from Intel Corporation. The views and conclusions herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of these organizations or the US government.

References:

1. Kozen, Dexter, "Language-Based Security", Computer Science Technical Report #TR99-1751, Department of Computer Science, Cornell University, June 15, 1999, <http://cs-tr.cs.cornell.edu:80/Dienst/UI/1.0/Display/ncstrl.cornell/TR99-1751>

2. Payette, Sandra and Carl Lagoze, "Flexible and Extensible Digital Object and Repository Architecture," ECDL98, Lecture Notes in Computer Science, Vol. 1513, Springer, 1998, (draft at: <http://www.cs.cornell.edu/payette/papers/ecdl98/fedora.html>)
3. Payette, Sandra, Christophe Blanchi, Carl Lagoze, and Edward Overly, "Interoperability for Digital Objects and Repositories: The Cornell/CNRI Experiments," *D-Lib Magazine*, May 1999. <http://www.dlib.org/dlib/may99/payette/05payette.html>
4. Schneider, Fred B., "Enforceable Security Policies," Computer Science Technical Report #TR98-1664, Department of Computer Science, Cornell University, July 24, 1999, <http://cs-tr.cs.cornell.edu:80/Dienst/UI/1.0/Display/ncstrl.cornell/TR98-1664>
5. Erlingsson, Ulfar and Fred B. Schneider, "SASI Enforcement of Security Policies: A Retrospective," Computer Science Technical Report #TR99-1758, Department of Computer Science, Cornell University, July 19, 1999, <http://cs-tr.cs.cornell.edu:80/Dienst/UI/1.0/Display/ncstrl.cornell/TR99-1758>
6. Erlingsson, Ulfar and Fred B. Schneider, "IRM Enforcement of Java Stack Inspection," Computer Science Technical Report #TR2000-1786, Department of Computer Science, Cornell University, February 19, 2000, <http://cs-tr.cs.cornell.edu:80/Dienst/UI/1.0/Display/ncstrl.cornell/TR2000-1786>.
7. *XrML driving the standard*, www.xrml.org
8. Blaze, Matt, Joan Feigenbaum, Martin Strauss, "Compliance Checking in the PolicyMaker Trust Management System", ATT Labs, 1998, <http://www.research.att.com/library/trs/TRs/98/98.3/98.3.2.body.ps>.
9. *IBM Cryptolopes*, <http://www-4.ibm.com/software/security/cryptolope/>
10. *PRISM Home Page*, <http://prism.cornell.edu>
11. Nelson, M., Maly, K., Zubair, M., and Shen, S.N.T, SODA: Smart Objects, Dumb Archives, ECDL99, Lecture Notes in Computer Science, Springer, 1999, <http://link.springer.de/link/service/series/0558/bibs/1696/16960453.htm>
12. *The KeyNote Trust Management System*, <http://www.crypto.com/trustmgt/kn.html>

INDIGO — An Approach to Infrastructures for Digital Libraries

Christian Mönch

Fachbereich Informatik, Johann Wolfgang Goethe-Universität,
D-60054 Frankfurt, Germany,
moench@tm.informatik.uni-frankfurt.de

Abstract. In this paper INDIGO, an approach to infrastructures for digital libraries is presented. It fulfills two crucial requirements to digital libraries: scalability and the ability to handle newly evolving document types. Based on a classification of digital library architectures, the main reasons for limited scalability and extensibility of digital libraries are identified. To overcome the identified problems the concept of mobile structure knowledge, on which INDIGO is based, is developed. The architecture of INDIGO is outlined and examples for the application of the concept are given.

1 Introduction

Digital libraries are long living applications that are capable of storing, processing and presenting a large number of electronic documents, that may vary widely in type and structure, e. g. text documents, audio/video documents, and hypermedia documents. There are three crucial requirements to digital libraries, they should be *orthogonal*, *scalable*, and *extensible*.

A digital library is orthogonal if it provides a set of operations, so called *orthogonal operations*, that can be applied transparently to every document in the library. Only the availability of orthogonal operations enables the handling of large numbers of different documents and allows for automated processing of documents, e. g. search (cf. [18,14]). It is beyond the scope of this paper to define a complete set of orthogonal operations, but we suggest some operations, e. g. Present, that might be part of such a set.

A digital library is scalable if it allows for a virtually unlimited number of documents and clients. This requirement leads to an implementation of digital libraries as decentralized, distributed systems (cf. [1]).

A digital library is extensible if it allows the integration of newly developed document types. Integration refers to the ability to execute the orthogonal operations on those documents. Extensibility enables digital libraries to cope with newly evolving document types, that will evolve during the lifetime of the digital library (cf. [11]). Examples for the fast pace of development are HTML, VRML, HyTime, and XML. The availability of meta languages such as XML even increases the invention rate, enabling the definition of formats like SVG, SMIL, MathML, and X3D.

In the next section we will argue why it is difficult to simultaneously implement scalability and extensibility in orthogonal digital libraries.

2 Document Processing in Digital Libraries

To understand why digital library architectures that support orthogonal operations provide only limited support of simultaneous scalability and extensibility, we have to take a closer look on document processing in orthogonal digital libraries.

2.1 Performing Orthogonal Operations

In an orthogonal digital library users request the execution of orthogonal operations on documents. Every such request will eventually lead to the instantiation of a piece of code that performs the desired operation on the specified document. This code is further referred to as a *processor*. Processors interact with their environment in different ways. During their invocation they receive arguments from the caller, i. e. user clients to which they return a result. During their runtime they access resources and basic services of the environment, e. g. user-interface devices, secondary storage, and network transport services.

Some processors use the resources to convert and emit document data, for example, if they are involved in presentation or transportation of documents. Others do not emit data but calculate a result only in order to support further computation. Processors of the later class typically perform some kind of search operation on the document data. It should be noted that presentation of a document is only a side effect of the invocation of an appropriate processor.

Processors build the glue between orthogonal operations and different document types, by performing orthogonal operations in a document-specific way. For each document type and each orthogonal operation a distinct processor is needed, that possesses built-in knowledge about the structure of the document, so called *structure knowledge*. In order to be extensible, an orthogonal digital library must support the seamless integration of new structure knowledge.

2.2 Architectures of Digital Libraries

We assume that a digital library is a distributed application that consists of components that communicate over a network. Components are partitioned in servers and clients. Servers are components that store documents, clients are components that interact with an user. Given these assumptions, digital library architectures can be distinguished according to the presence and absence of structure knowledge in the client and in the server.

Structure Knowledge at Client Side. In this configuration, which is depicted in figure 1 a), processors are executed in the client. They use the basic services of the environment to communicate with the server and access the document data. This configuration is widely used today, e. g. in the WWW. It is also employed by different digital libraries, e. g. in the Informedia Project [17], and is the paradigm behind the architecture of the Kahn/Wilensky-repositories [7].

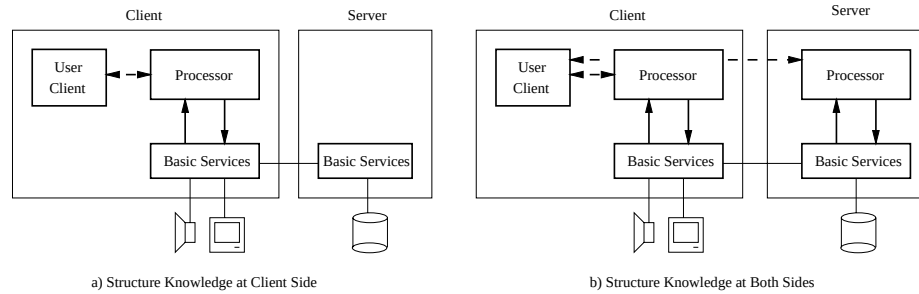


Fig. 1. Possible Distribution of Structure Knowledge

The advantage of the approach is its high scalability. Because servers need only provide remote access to their basic services. Because these are not document-specific, they can host every document type. The installation of new servers does therefore provide new resources for all document types. This approach is also well suited for presentation of documents, especially of interactive documents because presentation is performed locally on the client side.

A disadvantage of this architecture is the high bandwidth requirement for operations that perform calculations on a large number of documents, e. g. search, making it virtually impossible to perform these operations on a large number of documents. Another disadvantage is the limited support of heterogenous clients and updates of the client platforms, due to the necessary implementation effort for new processors and the porting effort for existing processors.

Structure Knowledge at Both Sides. In this configuration, depicted in figure 1 b), processors are executed in clients and servers simultaneously, enabling them to apply document specific protocols to exchange data. This model is often applied for streaming data, e. g. RealAudio, RealVideo, but it is also used to implement services with proprietary interfaces, e. g. WAIS, Z39.50, the QBIC image retrieval system [3], and WWW-Gateways¹.

The advantage of this approach is its support for an optimal placement of data-centric and user-centric methods, allowing to minimize bandwidth requirements. On the other side this approach limits the scalability of the digital library, because servers become typed by the built-in structure knowledge and can only be used for certain document types. As a result the number of supported documents does not only depend on the total number of servers, but on the number of servers of the matching type. In addition this configuration inherits the disadvantages of the previously described configuration, i. e. limited support for heterogenous clients and changes of client platforms.

Further Possible Configurations. Two possible configurations, i. e. structure knowledge at server side and structure knowledge at no side, have not been discussed so far. It

¹ WWW-Gateways seem to belong to the server-side model, but the results are presented as HTML-pages which may require a substantial amount of further computation.

is obvious that the later one does not support any orthogonal operation and will merely serve as transit systems.

Only few systems contain structure knowledge at server side exclusively, e. g. host-based systems with terminals at the client side, and X-clients using the X-server at the client side to interact with the user. The approach shows the disadvantages of the typed servers and presentation capabilities are strongly dependent on the available bandwidth. Therefore this configuration is found seldomly in wide area distributed systems, despite the fact that it supports efficient search processes and allows updates of clients.

2.3 Conclusion

The support for new document types requires the installation of new structure knowledge, preferably in the client and server components, otherwise presentation of documents or searching in documents may not be supported adequately. But the installation of structure knowledge at the server side limits the scalability of the number of documents by typing the servers. The installation of structure knowledge at the client limits the capability to include new client platforms or update existing client platforms, because of the resulting implementation and migration efforts. Alternatively the heterogeneity of structure knowledge could be limited to very few document types, but this would also limit the extensibility of the digital library.

In the remainder of the paper we will present INDIGO, our approach to an infrastructure for digital libraries, that are scalable while retaining extensibility.

3 INDIGO — A New Approach

To overcome the limitations that have been outlined in section 2, we developed an infrastructure for digital libraries (INDIGO). To cope with the limitations resulting from the static placement of structure knowledge the concept of *mobile structure knowledge* is introduced. The basic idea is that mobile structure knowledge can be moved to the place where it is needed, i. e. where an operation can be performed in the most efficient way. The following architectural elements are introduced in INDIGO: *document methods*, *meta documents*, and *execution servers*. They are shortly described in the following sections, for a more detailed description of the basic architecture please refer to [9].

3.1 Document Methods

A document method is a piece of mobile code that carries structure knowledge, i. e. a mobile processor. For every document and every mandatory operation a document method has to be provided.

The orthogonal set identified so far contains the following mandatory operations: Interface, Present, and Similar. The Interface-operation takes no argument and returns a list of operations that the document supports. The Present-operation performs the presentation of the document on the execution server given as argument. The Similar-operation calculates the similarity between the list of words given as arguments and the document. The similarity is returned as a fractional number between zero and one.

It is evident that the orthogonal set will never be completely specified. The necessity to introduce new operations will eventually occur. The flexibility of the operation mapping, allows for introduction of new orthogonal operations by agreeing about a new operation name. But most documents define privately used operations to refer to document specific operations (see example documents in section 4.1). To ensure that no conflicts between private operation names and names of newly introduced orthogonal operations occur, the names of private operations have to have the prefix *local*.

3.2 Meta Documents

To implement the concept of mobile structure knowledge, documents have to be associated with their document methods, i. e. the implicit mapping of orthogonal operations on document methods is made explicit. Documents in INDIGO consist of three parts, which are encoded in meta documents: a set of attributes, a mapping from orthogonal operations to document methods, and the content of the document.

The attributes store meta data about the contained document — not about the meta document itself — including the author of the document, the owner of the document, access rights, etc. The mapping specifies the mapping of orthogonal operations on document methods. Basically it is a mapping from names to URLs specifying the locations of the code of the corresponding method, and the required runtime environment. The document content may also be contained in the meta document. Optionally a meta document with empty content can be generated (see section 4.1 for an application of empty content).

A meta document with non-empty content can be interpreted as the definition of a class and its only instance — an object representing a document. The self-containment of the meta document prevents the need to keep a global repository, thus enabling the use of new document types, without registering a global type definition.

3.3 Execution Server

Document methods must be invoked in some way and eventually executed to perform the desired operations on the documents. These actions are supported by execution servers. Execution servers satisfy several needs, they store documents, accept commands to invoke orthogonal operations, and provide an environment for the execution of document methods. Every document that is stored by an execution server has an identifier that is unique in the context of the server. Together with the unique address of the server it forms a globally unique document identifier. Documents that are transported to or from an execution server are encoded as meta documents.

Server Commands. Execution servers offer several commands at their remotely accessible interface: Deliver, Invoke, Attributes, Delete, Documents, and Environments. Deliver allows to move a document to the execution server. Invoke supports the invocation of orthogonal operations on documents, Attributes retrieves the attributes of document and Delete removes a document from the server. The Documents-command returns the identifiers of the documents that are stored on the server, the Environments-command returns information about the runtime environments supported by the server.

Execution of Document Methods. The execution of an orthogonal operation on a document is performed in three steps. First the execution server consults the mapping to retrieve the appropriate document method and to determine the required runtime environment. In a second step the document method is moved to the execution server. Finally the runtime environment is created and the document method is executed.

The runtime environments provide mechanisms that enable document methods to access server resources, e.g. display and storage devices, in a controlled way. They also support a digital library interface (DL-interface), that enables document methods to execute server commands on local or remote servers and to deliver the document to another server.

3.4 Overall Architecture

In INDIGO a digital library is comprised by a number of loosely coupled execution servers. Every execution server has a set of *neighbours*, i.e. execution servers that are directly known to it. The closure on all neighbours comprises the server side of a digital library. Execution server on the client side are normally not contained in a neighbour set. Figure 2 illustrates the elements of INDIGO.

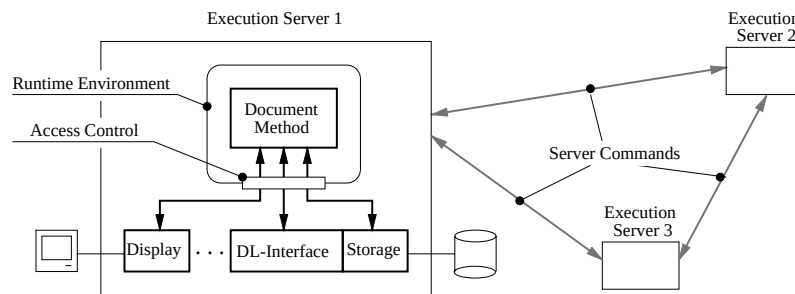


Fig. 2. Architectural Elements of INDIGO

4 INDIGO Prototype

To gain experiences with the concept, two prototypes were implemented, one in Python and one in Java. While the Python-based prototype is a straight forward implementation of the given concepts that supports methods written in Python and Java, the Java-based prototype has been designed to be easily integratable into existing infrastructures, i.e. to run as an applet in Java-enabled WWW-Browsers. To meet the requirements that hold for applets it does only support Java-based methods.

Currently only methods written in interpreted languages are supported in order to support mobility and security. Due to their uniform representation, interpreted methods are highly mobile. Interpreters can protect servers against malicious documents because they exercise fine-grained control about function calls by the documents method.

4.1 Examples

Offline Presentation. This example demonstrates the ability of the infrastructure to support new media types and data formats. The example document contains an image,

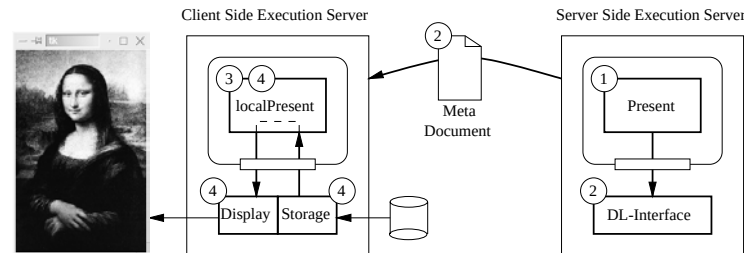


Fig. 3. Offline-Presentation of a Document

that shall be presented at the client side. This involves four processing steps. The presentation is initiated by an invocation of the Present-method of the document with the address of the client side execution server as argument. This results in the instantiation of the corresponding document method (step one in figure 3). The document method instructs the server, via the dl-interface, to move a copy of the document to the execution server at the client side (step two). After the document is copied to the client side, the Present-method invokes the localPresent-method on the copy and terminates (step three). The localPresent-method uses the system resources to access the document data and present the image to the user (step four).

It is obvious that there are no limitations for the data format of the document and only few limitations for the media type. New media types are supported as long as they can be mapped on existing presentation media such as text, image, audio, and video. However the presentation of documents containing, for example, olfactoric or haptic content may require new presentation devices.

Online Presentation. The second example demonstrates the ability of the infrastructure to support document specific distribution of structure knowledge. The example document contains a video stream that shall be presented at the client side. Again the presentation is initiated by invoking the Present-operation of the document, resulting in the instantiation of the corresponding document method (step one in figure 4). But the Present-operation of the video document does not move a complete copy of itself to the client side execution server. Instead it moves a copy with empty content (cf. section 3.2) to the client side execution server (step two). Afterwards the Present-method creates an RTP-socket and invokes the localPresent-Method on the client side with the URL of the RTP-socket as argument (step three). The localPresent-method uses the RTP-socket to receive the document data from the Present-method and presents it to the user (step four). The video document places its structure knowledge at the client and at the server side to support an efficient presentation. It was implemented using the Java Media Framework (JMF).

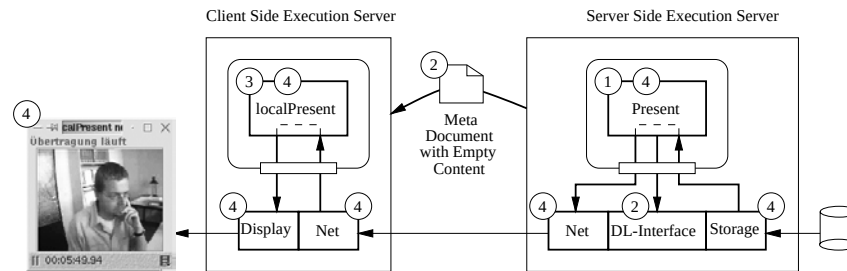


Fig. 4. Online-Presentation of a Document

Our experiments have shown that audio and video streaming is currently at the edge of performance in interpreted systems. It support transport and presentation of low quality video and audio streams only.

Distributed Searching. In our infrastructure a digital library is comprised of a set of loosely coupled execution servers, only related by their neighbour-sets. There is no central instance, i. e. no instance that stores the state of all servers. It is a non-trivial task to implement library services, e. g. catalog and search aids, in such an environment. The implementation of such services requires the ability to perform operations, e. g. querying of document list or searching through local documents, on all servers. This involves two tasks: enumerating all servers and invoking operations on each server, perhaps on each document in the server.

To find all servers we implemented an algorithm that performs a distributed calculation of a spanning tree between the servers. Such an algorithm could be implemented using documents that flood the neighbour set and leave footprints on each server in order to recognize cycles. But in this approach extra care must be taken to remove the footprints. We decided to let the execution servers perform the spanning tree calculation using well-known distributed algorithms and store the result in a document with the special name *spanningtree* that is present at every server.

Given these preconditions, we implemented a document that uses the Similar-operation to search for the 100 documents in the library that are most similar to a given set of keywords. Besides the mandatory operations the document contains an operation named *localStart*. To initiate the search the *localStart*-operation is called with a set of keywords and an optional argument indicating the source server. The *localStart*-operation uses the Documents-command of the server to determine all locally stored documents and invoke the Similar-operation on every document, recording the 100 documents that are most similar to the given keywords. Thereafter it moves a copy, further referred to as *sub-document*, of itself to every execution server listed in the *spanningtree*-document of the local server, except the source server. It invokes the *localStart*-operation on every sub-document, giving the keywords and the local server as argument. The results of the method executions, 100 document-identifiers with their similarity measure, are then integrated into the local result list. After each sub-document terminated, the overall result is returned to the caller.

5 Related Work

The concept to encapsulate structure knowledge in mobile entities has been applied for Uniform Resource Agents (URAs) [2] that are used in the Silk environment to mediate between resources and users. It has been described by the author in [9]. An early approach that allows the transportation of structure knowledge to clients, in order to extent presentation and processing capabilities, are Multivalent Documents described in [13]. The need to incorporate intelligent behaviour into documents has also been recognized in the SODA-approach [10] and in the FEDORA-project [12]. However, both approaches do not provide an orthogonal set of operations. Similar ideas were developed in the field of distributed applications in the Globe-architecture [16]. In the AspectIX infrastructure [5], interfaces are separated from their implementation and implementations can be dynamically exchanged to match the desired behaviour. Mobile structure knowledge is also used in the context of multimedia databases [15], in Active Networks, and in the JINI architecture.

A large number of projects deals with mobile code and mobile agents. Mobile agent systems like TACOMA [6], and ffMain [8] show different ways to build mobile code and mobile agent systems. Some of the ideas given there are used in INDIGO.

6 Conclusion and Future Work

In this paper INDIGO was presented, an approach to architectures for digital libraries. It allows for the seamless integration of new document types into digital libraries by using mobile structure knowledge, thus avoiding most of the disadvantages that static placement of structure knowledge exhibits. New document types can be introduced without any global registration, thus providing authors with the means to design innovative documents that can be instantly used.

This flexibility is achieved by the installation of execution servers on every host that should become part of the digital library. Beside their main purpose, to support mobile structure knowledge, execution server also provide a homogeneous environment helping to overcome scalability limitations that stem from heterogenous platforms. A drawback of the approach is, that these controlled environments are based on interpretation of programmes, where performance becomes an issue. Slim binaries [4], that are translated into the native code of the host machine may be employed, if the security problems were solved and if support for platform-independent resource access was available.

Some aspects of the infrastructure, that are currently investigated, have not been discussed in this paper, e. g. security issues and native code-methods for well known document types. Future work will include investigation in the implementation of distributed algorithms in INDIGO as well as support for media-independent content-based search.

Acknowledgements

The author thanks Martin Heß and Prof. Dr. Oswald Drobnik for their comments and for fruitful discussions. This work was partially funded by a grant from the “Deutsche Forschungsgemeinschaft” in project BIB45 FRuv 01–02.

References

1. William P. Birmingham. An Agent-Based Architecture for Digital Libraries. *D-Lib magazine*, July 1995. [hdl:cnri.dlib/july95-birmingham](http://hdl.cnri.dlib/july95-birmingham).
2. L. Daigle, P. Deutsch, B. Heelan, C. Alpaugh, and M. Maclachlan. Uniform Resource Agents (URAs). [ftp://ftp.denic.de/pub/rfc/rfc2016.txt](http://ftp.denic.de/pub/rfc/rfc2016.txt) [7.12.1999], 1996. RFC 2016.
3. Marion Flickner, Harpreet Sawhney, Wayne Niblack, Jonathan Ashley, Qian Huang, Byron Dom, Monika Gorkani, Jim Hafner, Denis Lee, Dragutin Petkovic, David Steele, and Peter Yanker. Query by Image and Video Content: The QBIC System. *IEEE Computer*, Vol.28(Nr.9), September 1995.
4. Michael Franz and Thomas Kistler. Slim binaries. *Communications of the ACM*, 40(12):87–94, December 1997.
5. Martin Geier, Martin Steckermeier, Ulrich Becker, Franz J. Hauck, Erich Meier, and Uwe Rasthofer. Support for mobility and replication in the AspectIX architecture. Technical Report TR-14-08-05, IMMD IV, Universität Erlangen-Nürnberg, 1998.
6. Dag Johansen, Robbert van Renesse, and Fred B. Schneider. An Introduction to the TACOMA Distributed System Version 1.0. Technical Report 95–23, June 1995.
7. Robert Kahn and Robert Wilensky. A Framework for Distributed Digital Object Services. [hdl:cnri.dlib/tn95-01](http://hdl.cnri.dlib/tn95-01), 1995.
8. A. Lingnau, P. Dömel, and O. Drobniak. A HTTP-based Infrastructure for Mobile Agents. In *Proceedings of the 4th International World Wide Web Conference*, Boston, 1995.
9. Christian Mönch and Oswald Drobniak. Integrating new document types into digital libraries. In *Proceedings of the IEEE Forum on Research and Technology Advances in Digital Libraries IEEE ADL '98*, pages 56–65. IEEE, April 1998.
10. Michael L. Nelson, Kurt Maly, Mohammad Zubair, and Stewart N. T. Shen. Soda: Smart objects, dumb archives. In *Research and Advanced Technology for Digital Libraries — Third European Conference, ECDL'99*, volume 1696 of LNCS, pages 453–464, Paris, Frankreich, September 1999.
11. Peter J. Nürnberg, Richard Furuta, John L. Leggett, Catherine C. Marshall, and Frank M. Shipmann III. Digital Libraries: Issues and Architectures. In *Proceedings of The Second International Conference on the Theory and Practice of Digital Libraries*, pages 147–153, June 1995.
12. Sandra Payette and Carl Lagoze. Flexible and Extensible Digital Object Repository Architecture (FEDORA). In Christos Nijolaou and Constantine Stephanidis, editors, *Research and Advanced Technology for Digital Libraries*, volume 1513 of LNCS, pages 41–59. Springer, 1998.
13. Thomas A. Phelps. *Multivalent Documents: Anytime, Anywhere, Any Type, Every Way User-Improvable Digital Documents and Systems*. PhD thesis, University of California, Berkeley, 1998.
14. Bruce Schatz and Hsinchun Chen. Building Large-Scale Digital Libraries. *IEEE Computer*, 29(5):22–26, May 1996.
15. Val Tannen. Heterogenous data integration with mobile information managers. In Sushil Jajodia, M. Tamer Özsu, and Asuman Dogac, editors, *Proceedings of Advances in Multimedia Information Systems, MIS'98*, volume 1508 of LNCS, pages 2–3. Springer, 1998.
16. Maarten van Steen, Philip Homburg, and Andrew S. Tanenbaum. Globe: A wide-area distributed system. *IEEE Concurrency*, pages 70–78, Jan–Mar 1999.
17. Howard W. Wactlar, Takeo Kanade, and Scott M. Stevens Michael A. Smith. Intelligent Access to Digital Video: Informedia Project. *IEEE Computer*, 29(5):46–47, May 1996.
18. Michael P. Wellman, Edmund H. Durfee, and William P. Birmingham. The digital library as a community of information agents. *IEEE Expert – Intelligent Systems & Their Application*, 11(3):10–11, June 1996.

Scalable Digital Libraries Based on NCSTRL/Dienst

Kurt Maly, Mohammad Zubair, Hesham Anan, Dun Tan, and Yunchuan Zhang

Department of Computer Science
Old Dominion University, Norfolk, VA 23529, USA
{maly, zubair, anan, tan_d, yczhang}@cs.odu.edu

Abstract. NCSTRL (The Networked Computer Science Technical Report Library) is a successful digital library for scientific and technical information. It uses the Dienst protocol that was developed by ARPA-funded CS-TR project. We encountered several problems while implementing NCSTRL based large-scale libraries: UPS for Los Alamos and JDL for JTASC. The document collection for these libraries can range from several hundred thousands to few millions. The first problem we found that the native Dienst implementation does not scale beyond approximately 30,000 records. Secondly we found that the implementation is tightly coupled to the Unix platform. Finally, for a large number of hits the NCSTRL search interface support is limited in terms of usability. To address these problems, we replaced the Dienst repository service implementation with an Oracle-based implementation using servlet technology. The Oracle database stores the index information (metadata) and is partitioned horizontally to speed searching through different archives. Furthermore, indexes were built in order to speed the search by different key items such as the author name, the title and the abstract. Our implementation significantly reduced the average wait time for a user for searches that resulted in a large number of hits. In addition, we get all the other benefits of using servlet technology such as efficiency and portability. In this paper, we present the performance results of the new implementation and compare it with that of the implementation of the Dienst protocol in NCSTRL.

Introduction

Digital libraries (DLs) is an important research topic in many scientific communities and has already become an integral part of the research process. Currently, there are a number of commercial products available for individual communities to create their specialized digital library (for example, <http://www.software.ibm.com/is/dig-lib/v2factsheet>). Similarly, the research community has created excellent production digital libraries systems: NCSTRL/Dienst [3], the Digital Library Initiative [15], the Physics e-Print Server, RIB (Netlib/NHSE) [2], TRSkit [12], the ADS [1], CORE [6], etc. NCSTRL (Networked Computer Science Technical Report Library) is one of the first successful, operational, distributed DLs on a national scale. It is a library of computer science technical reports that covers a significant portion of the current CS research. Approximately 100 publishing organizations in 3 continents currently make up this collection, and this number is continuously growing. NCSTRL [10] consists of

a body of about 20,000 reports, high quality standards, and it contains a management process for publishing into the library.

The Dienst digital library protocol [3] is the engine used to support both NCSTRL and a related project, NCSTRL+ [13]. The current implementation of NCSTRL is not suitable for large-scale libraries. We encountered several problems while implementing NCSTRL based large-scale libraries: Universal Preprint Service (UPS) for Los Alamos National Laboratories and the Joint Digital Library (JDL) for JTASC. The UPS library was constructed as a proof-of-concept of a multi-discipline digital library that harvested its nearly 200,000 records from several different archives. The UPS prototype was used as the discussion piece for the UPS meeting, held in Santa Fe, New Mexico, October 1999, sponsored by Los Alamos National Laboratory that resulted in the Santa Fe Convention and the forming of the OpenArchiveInitiative [17]. The Joint Digital Library (JDL) for JTASC is being built to harvest documents on the order of millions from different defense related sites. Besides harvesting, in the JTASC project we are also addressing the issue of auto-classification scheme for a set of defense related sites. Both these libraries use bucket technology [13] to represent digital objects. Buckets are digital object that can be aggregated into complex objects and they have a communication protocol that allows them to behave intelligently.

The document collection for these libraries can range from several hundred thousands to few millions. The first problem we found that the Dienst implementation used by NCSTRL does not scale beyond approximately 30,000 records. Secondly we found that the Dienst software architecture tightly coupled to the Unix platform. Finally, for a large number of hits the NCSTRL search interface support is limited in terms of usability. To address these problems, we replaced the Dienst repository service implementation with an Oracle-based implementation using servlet technology. The Oracle database stores the index information (metadata) and is partitioned horizontally to speed searching through different archives. Furthermore, indexes were built in order to speed the search by different key items such as the author name, the title and the abstract. Our implementation significantly reduced the average wait time for a user for searches that resulted in a large number of hits. In addition, we get all the other benefits of using servlet technology such as efficiency and portability [8]. Once the servlet is invoked, it remains in the server's memory and separate lightweight threads handle multiple concurrent requests. This is unlike the existing CGI Perl script implementation where every request results in a separate process and invocation of the Perl interpreter. This makes servlets more efficient and scalable. The servlets are portable across operating system and across server implementations. This is because servlets are written in Java and follow a widely accepted API. In this paper, we present the performance results of the new implementation and compare it with the existing NCSTRL implementation.

The rest of this paper is organized as follows: section 2, discusses NCSTRL implementation and its limitations, section 3 describes our enhancement to NCSTRL and NCSTRL+; section 4 discusses the experiments performed and discusses the performance analysis results. Finally, section 5 concludes the paper and presents future work.

NCSTRL and Its Limitations

In this section, we discuss NCSTRL and its implementation limitations. Fig 1 shows the architecture of NCSTRL/Dienst [3]. Its repository service is based on the Unix file system and this introduces scaling problems:

- A conflict between the Dienst implementation and the Solaris operating system occurs. Each Dienst publishing authority expects all of its publications to be in a single directory. Even though the UPS and JTASC DLs use buckets, the buckets are internally stored and indexed in the same manner as regular Dienst objects. Thus, each bucket occupies 1 subdirectory in the Dienst publishing authority, and Dienst requires that all documents be at the same level within the publishing authority. The Solaris operating system has a limit of 32,767 subdirectories within a directory. As such, for the large archives, it is not possible to have all the publications in a single publishing authority.
- Problems with the search engine occur. The native Dienst search engine does not scale beyond approximately 20,000 records. The Dienst developers recommend the use of freeWAIS-sf for larger collections. Unfortunately, freeWAIS-sf has its own scaling problem, since it automatically considers a word to be a stop word if it has more than 20,000 occurrences. With the large number of records to be indexed, many non-stopwords occur more than 20,000 times (e.g. “galaxies”, “dimensional”, “temperature”, “spin”, etc.). More importantly, the archive names themselves become stopwords, making browsing through cluster selection impossible. In addition, the freeWAIS-sf only returns a partial listing of hits –the ones it determines to be most relevant. If 1,000 records match a search, freeWAIS-sf will return less than 200. This effectively negates the ability to do browsing on clusters and advanced searching, an important feature in NCSTRL+ the basis for both UPS and JTASC digital library.

An Efficient Dienst Implementation

So far we have identified the problems of currently existing Dienst implementations. We set out to create a realization that will provide a short response time, for large number of concurrent users searching a large number of records. In Figure 2 we illustrate the information retrieval process based on our design of the Dienst protocol implementation – compare this to Figure 1.

- A client submits a query (through web browser) to the search server.
- The search servlet on the server processes this query and retrieves the results from an Oracle database (which contains the metadata).
- The search servlet passes a subset of the results to the client and information to be used later in getting the rest of the results.

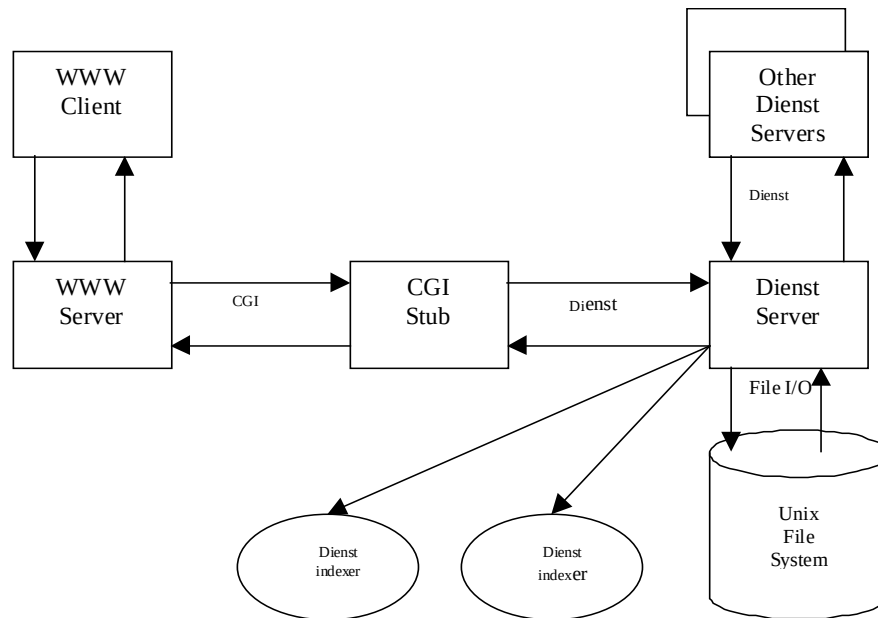


Fig 1. NCSTRL/ Dienst Architecture

In the remainder of this section we describe how the Oracle Database is designed and partitioned and the alternatives that we considered in designing the search server.

Database Schema

The testbed for our implementation is the UPS prototype of the OpenArchiveInitiative [16]. It contains six archives: CogPrints, NCSTRL, RePEc, xxx, NDLTD and NTRS. Although the same attributes (such as the author name, title, abstract, ...) were maintained for all archives, we decided not to store all the metadata in a single table and partition the database horizontally by dedicating a separate table for each archive. This allows us to support browsing of individual archives without having to search the entire collection and we expect this to be a fairly common operation. In addition, our database only stores the metadata of the documents. The actual documents themselves are distributed cross the different archives.

Search Server Architecture

The search server is implemented in Java using Servlets. The components of the search server are shown in Figure 3. The session manager maintains one session per user per query. It is responsible for creating new sessions for new queries (or for queries whose session has expired). Sessions had to be used because queries are

expected to return large number of results (hits) that can't be displayed on one page. Thus sessions are used to cache some or all the results (as will be explained later) in order to make browsing through the hits faster. The session manager receives two types of requests from the client. The session manager receives either a request to process a new query (search), or a request to retrieve another page of results for a previously submitted query (browsing). For a search request, the session manager

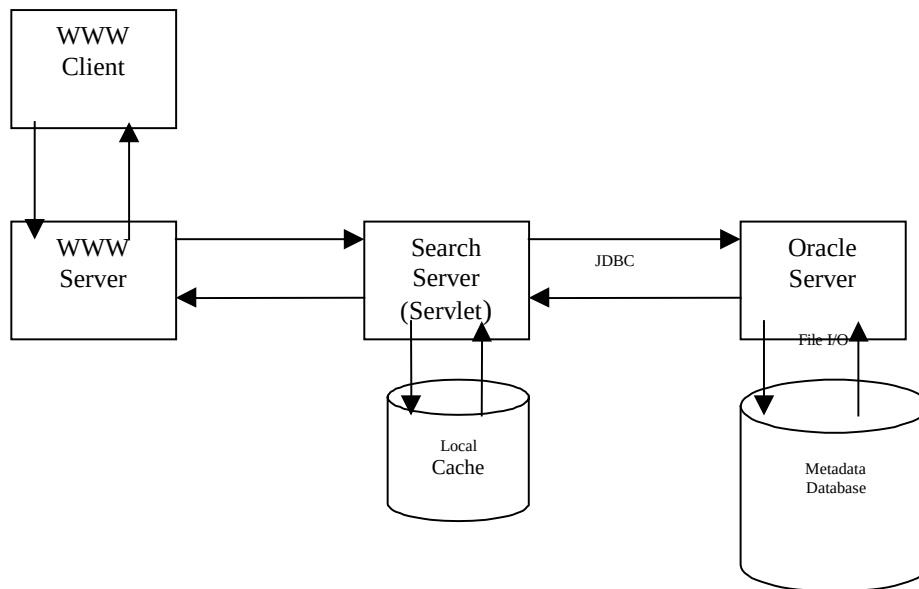


Fig 2. Information retrieval process based on Oracle implementation

calls the index searcher that formulates a query (based on the search parameters) and submits it to the Oracle server (using JDBC) then retrieves the search results. The session manager then calls the result displayer to display the first page. For a browsing request, the session manager checks the existence of a previous session (session are expired after a specific time of inactivity).

If a previous session is not found (expired), a new session is created and the search is re-executed, then the required page is displayed. In the case where the previous session still exists, the required page is displayed based on the cached data (which may require additional access the Oracle database). The choice of which data to be cached is illustrated in details in the following subsections.

Full Results Caching. In this approach, once the query is executed, all the hits should be retrieved and cached in memory to be used later in page display. This approach was not implemented because the number of search results is expected to be very large and caching all of them will consume large amount of memory resources.

Full / Partial Results Caching. In this approach, once the query is executed, part of the results are retrieved in full and cached (in the full result set). For the rest of the results, only ids (database keys) are retrieved and cached (in the partial result set). Displaying results for hits that exist in the full list is trivial but displaying results for hits that exist in the partial list will require additional access to the Oracle database to get the full results.

This approach will enhance the browsing time per page. However, the response time for the end user will be sluggish because the user will have to wait for all the results (or their Ids) to be retrieved.

Full / Incomplete Partial Caching. This approach is similar to the previous one but instead of caching the complete set of results only part of them is cached. Upon request to retrieve un-cached results, the search will be re-executed and another set of results will be cached. This will enhance the response time. However some pages might take longer time to display.

Incomplete Partial Caching. Experiments have shown that the difference between displaying time for a page whose full data are cached locally and displaying time for a page where only Ids are cached is not significant. So it has been decided to cache only part of Ids of the results (incomplete partial list).

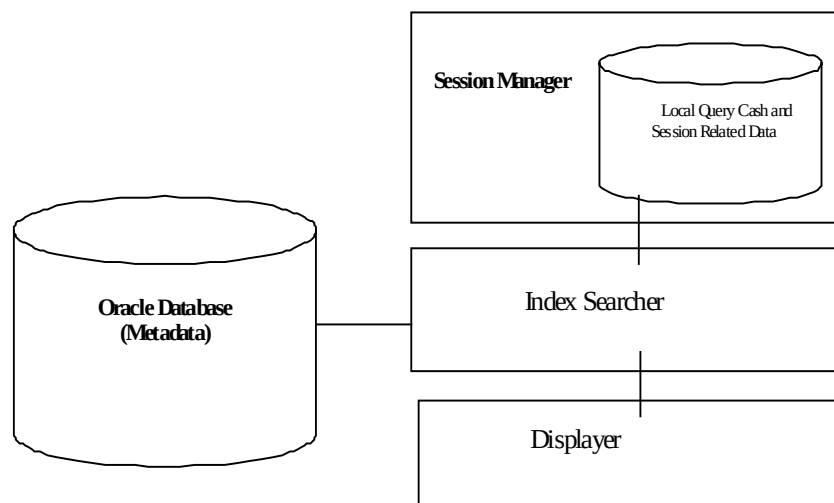


Fig 3. Search Server

Experimental Results

In this section, we present our performance results, evaluating and comparing the performance of traditional NCSTRL searches and the new efficient JTASC implementation. For user satisfaction we take the traditional ‘response’ metric as well as a new metric we call ‘browse’ measure. We also compare the number of hits returned by traditional Dienst implementation and our JTASC implementation.

Some queries were selected for the experiments and they are shown in table 1.

Search #	Search Query
1	author=David, archives=CogPrints
2	author=Maly
3	title=communication
4	author=John, archive=NCSTRL
5	title=network
6	author=John
7	author=David

Table 1. Queries

Following is the test-bed we used: the Dienst is running on a Sun Ultra 5 with 256M RAM and SunOS5.6; our servlet version implementation is running on a Sun Ultra 5 with 256M RAM and SunOS5.7. All simulated users are from a NT machine (Pentium III 450 with 128M RAM) with JDK1.3. All of these machines are in one same subnet.

Response Time

The response time is defined to be the time elapsed between the submission of the search query and the display of first page in the results. This time includes the time to formulate a database query, execute this query and retrieve the results (either all the results or part of them depending on the caching policy). Figure 4 shows the average response time for different sets of queries when applied to the Full/Partial caching policy and the Incomplete Partial policy with size of cache ranging from 50 to 250.

Fig 4 shows that when the partial cache size increases, the response time increases slightly because more partial results have to be retrieved. For example the query the returns 887 hits has response time of 4.1 sec for a partial size of 50, 5.1 sec for a partial size of 100, 9.013 sec for a partial size of 200 and 16.03 sec for a partial size of 250. However the same query takes about 39 sec in the case of Full/Partial caching approach. This is due to the fact that the Full/Partial caching system should first execute the query then caches all the 887 hits (either fully or partially). This significant increase in the response time was the reason that we decided not to use this policy.

Also, as the number of hits increases, the response time increases (although we retrieve only part of the results). This is due to the internal processing of queries by Oracle that tries to create an internal table that contains the results of the submitted query. The figure shows some anomalies in the response time results. For example, the last query that has 4194 hits has response time less than the previous one that has 3833 hits. This is due to the fact that in some queries we either search all archives or only a single archive based on the client's choice. And obviously, the time for searching all archives will be somewhat larger than the time to search only one archive because in this case all the Oracle database tables will have to be accessed.

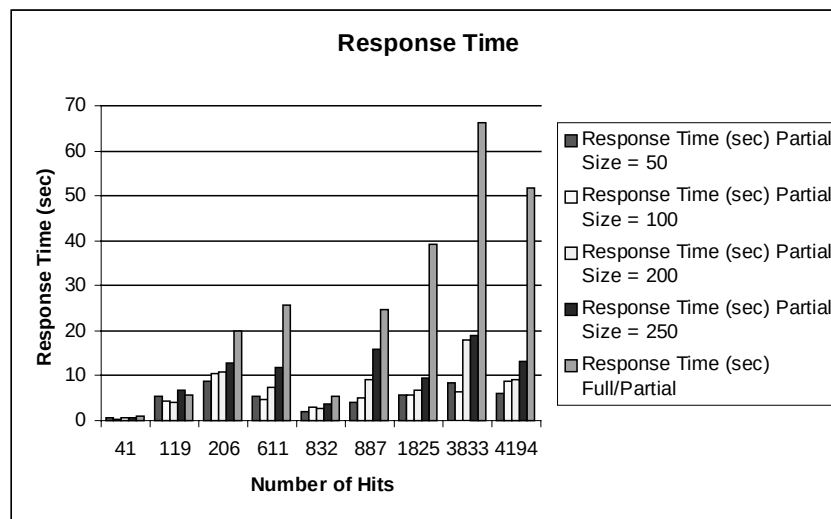


Fig 4. Effect of caching policy on response time

Fig 5 shows a comparison between the response time of Dienst and JTASC implementation. It is obvious from the graph that the response time for Dienst is less than the response time for JTASC implementation. However the difference in timing is not very significant. In other words the response time of JTASC implementation is comparable to the response time of Dienst. On the other hand, the number of hits that are returned by the JTASC implementation is in most cases larger than the number of hits returned by Dienst as will be illustrated in section 4.3.

Browsing Time

The browsing time is defined to be the average display time for a page (a page contains a maximum of 20 hits). This average does not include the first page time. Fig 6 shows the average display time per page (which is a measure of the browsing time). By browsing here, we mean the capability to view the results of search queries and navigate through them on page by page basis, which makes displaying the search results more convenient to the end user.

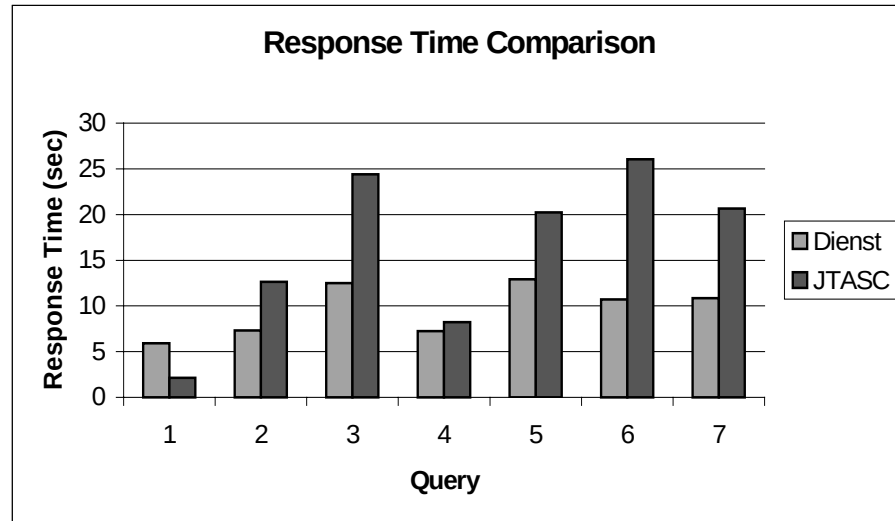


Fig 5. Response Time Comparison

Although the response time tends to increase when more partial results have to be retrieved, average page time tends to decrease because more results are cached and retrieving them later (browsing) will be faster.

For example the query that returns 1825 hits has average page time of 1.14 sec for a partial size of 50, 1.0023 sec for a partial size of 100, 0.9388 sec for a partial size of 200 and 0.9088 sec for a partial size of 250.

In conclusion, figure 4 shows that the response time for the Full/Partial caching policy is higher than that for the Incomplete partial approach. This is expected because in the Full/Partial approach, all results are either retrieved in full or partially cached. However the average page display time in this case is less than that for the incomplete partial approach. Based on the importance of response time (for interactivity), the Full/Partial approach was rejected and we decided to use the incomplete partial caching policy. This will degrade the average page display time by about 0.7 sec, which will make the browsing slightly slower but still not annoying.

Number of Hits

In JTASC implementation, all records that match the query elements - and only those records - are returned. This feature of the JTASC implementation allows us to use it for browsing. On the other hand in Dienst implementation, the number of hits returned by WAIS engine in the NCSTRL implementation is limited by the buffer size (it returns approximately no more than 200 hits).¹

Figure 7 shows a comparison between number of hits returned by Dienst and number of hits returned by JTASC implementation for the same queries. From the

¹ We contacted the original authors of WAIS who indicated that the size of the search result set is fixed somewhere deep in the code and not easily changed.

graph it is obvious that JTASC implementation returns more hits than Dienst. For example, query 7 returns 191 hits using Dienst implementation and 4194 hits using JTASC implementation.

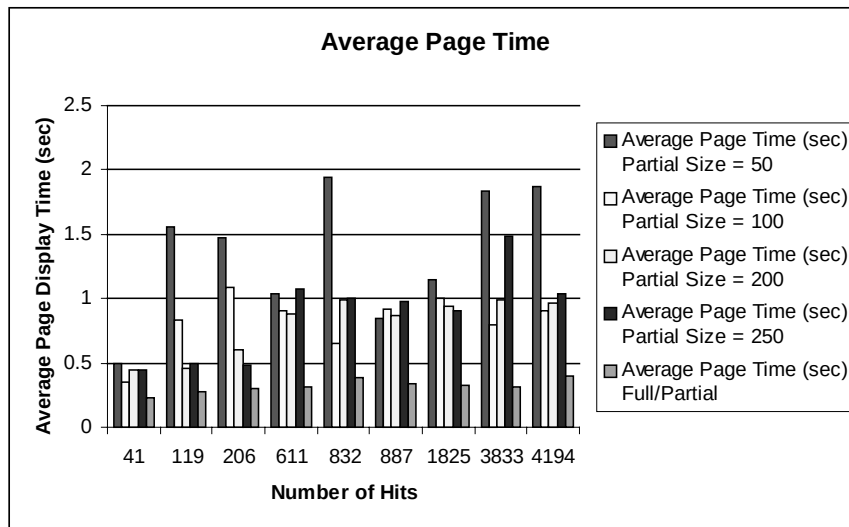


Fig 6. Effect of caching policy on average page display time

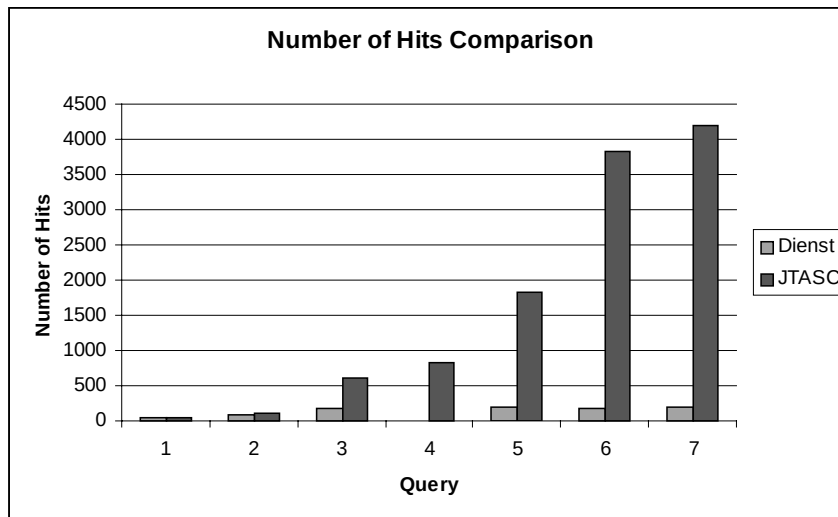


Fig 7. Number of Hits Comparison

Conclusions and Future Work

Dienst has been used very successfully in the NCSTRL distributed digital library and for the Santa Fe Convention of the OpenArchiveInitiative [7]. The NCSTRL implementation of Dienst is based on open source and freely available software: one, its repository service is built on top of the Unix file system, two its index service is built on top of WAIS, and three, its server uses CGI technology. All this contributes to a serious scaling problem in terms of number of concurrent accesses, number of documents in its repository, and three the number of hits returned. NCSTRL has basically reached its limit with about 30,000 documents and the current implementation cannot be used for realistic, large collections such as the UPS federation that contains 200,000 objects. We have re-implemented Dienst using free servlet technology and a commercial index and storage technology – Oracle. In this paper we have shown that our JTASC implementation solves some of these problems and that performance is well within traditional expectations of Web search engines, a metric that all digital libraries need to meet. Moreover, JTASC implementation is designed to be independent of the platform since it is based on Java technology.²

The search time of multiple tables (all archives) can be enhanced by submitting the query in parallel using separate thread for each table. This might significantly reduce the response time but it will require more complex session management strategy to be implemented.

Also, oracle has the distributed database support available. Several databases on different machines can be linked together through the network and serve the application's need. But due to the limitation of our current experimental environment, we do not implement this distributed architecture. It should be considered in our future work.

The concurrency response of this system can be enhanced if the search service is available on a set of machines instead of one. In this case we can add another service to balance the load on the search servers. This service will act as a gateway that takes the user query and pass it to the appropriate (the least loaded) search server then the client can interact directly with this service. Another service that can enhance the concurrent user response is the replication of the Oracle metadata; the trade-off for such a service is the large size of the data to be replicated. However, multiple processors can be added under supervision of the Oracle server and this will enhance the execution time for the queries.

On a longer time scale we plan to study the scalability limits of this approach particularly as it concerns the creation of federation of metadata collections to provide portals to large number of existing digital libraries.

References

1. Accomazi, A., Eichhorn, G., Kurtz, M. J., Grant, C. S., Murray, S. S.,: Astronomical information Discovery and Access: Design and Implementation of the ADS Bibliographic Services. *Astronomical data Analysis Software and Systems VI*, 125, (1997), 357-360

² We successfully ran the JTASC implementation on Unix and Windows NT platforms.

2. Browne, S., Dongarra, J., Horner, J., McMahan, P., Wells, S.: National HPCC Software Exchange (NHSE): Uniting the High Performance Computing and Communications Community. D-Lib Magazine, May (1998), <http://www.dlib.org/dlib/may98/browne/05browne.html>
3. Davis, J.R., Kraft, D. B., Lagoze, C.: Dienst: Building a Production Technical Report Server. *Advances in Digital Libraries*. Springer-Verlag, (1995), 211-222
4. Davis, J. R., Lagoze, C.: The Networked Computer Science Technical Report Library. Cornell CS TR96-1595, July, (1996)
5. Dushay, N., French, J. C., Lagoze, C.: A Characterization Study of NCSTRL Distributed Searching. Cornell CS TR99-1725, January (1999)
6. Entlich, R., Garson, L., Lesk, M., Normore, L., Olsen, J., Weibel, S.: Making a Digital Library: The Contents of the CORE Project. *ACM Transactions on Information Systems*, 15(2), (1997), 103-123
7. Ginsprag::Open Archives Initiative, <http://www.openarchives.org>
8. Hunter, J, and Crawfor, W.: Java Servlet Programming. O'Reilly and Associates, October 1998.
9. Leiner, B.M.: The NCSTRL Approach to Open Architecture for the Confederated Digital Library. D-Lib Magazine, December (1998)
10. Maly, K., French, J., Fox, E., Salman, A.: Wide Area Technical Report Service – Technical Reports Online. *Communications of the ACM*, p. 45, April (1995)
11. Maly, K., Nelson, M. L., Zubair, M.: Smart Objects, Dumb Archives A User-Centric, Layered Digital Library Framework. D-Lib Magazine, March (1999), Volume 5 Issue 3
12. Maly, K., Nelson, M. L., Shen, S. N. T., Zubair M.: Buckets: Aggregative, Intelligent Agents for Publishing. *WebNet Journal*, Vol. 1, No. 1, March (1999), 58-65
13. Nelson, M. L., Maly, K., Shen, S. N. T., Zubair, M.: NCSTRL+: Adding Multi-Discipline and Multi-Genre Support to the Dienst Protocol Using Clusters and Buckets. *Proceeding of Advances in Digital Libraries 98*, Santa Barbara, CA, April 22-24, (1998)
14. Schatz, B., Chen, H.: Building Large-Scale Digital Libraries. *IEEE Computer*, 29(5), (1996), 22-26
15. Schatz, B., Mischo, W. H., Cole, T. W., Hardin, J. B., Bishop, A. P., Chen, H.: Federating Diverse Collections of Scientific Literature. *IEEE Computer*, 29(5), (1996), 28-36
16. Sompel, H. V., Nelson, M.L., Lyapunov, V .M. , Zubair, M., Liu, X., Krichel, T., Hochestenbach, P., Maly, K., Kholief, M., O'Connell, H.:The UPS Prototype project: exploring the obstacles in creating a cross e-print archive end-user service. D-Lib Magazine, February (2000) , Volume 6 Issue 2
17. Sompel, H. V., Lagoze, C.: The Santa Fe Convention of the Open Archives Initiative. D-Lib Magazine, February (2000)

OMNIS/2: A Multimedia Meta System for Existing Digital Libraries

Günther Specht and Michael G. Bauer

Institut für Informatik
Technische Universität München
Orleansstraße 34, D-81667 München, Germany
{specht,bauermi}@in.tum.de

Abstract. Since today more and more complementary information is available in different electronic media there is an increasing demand for the integration of traditional digital library systems and multimedia systems. In this paper we present the OMNIS/2 system, which is an advanced meta system and enhances existing digital library systems or retrieval systems by additional storing and indexing of user-defined multimedia documents, automatic and personal linking concepts, annotations, filtering and personalization.¹ The key concept of OMNIS/2 is that all of the above mentioned features are accomplished without changing the underlying documents. In our architecture existing digital library systems, which are established applications, serve as a document storage layer, while OMNIS/2 forms the multimedia storage layer, linking layer and personalization layer. This general approach ensures the integration and transparent combination of different digital library systems. Thus with OMNIS/2, even mere retrieval systems - and nowadays most digital library systems are mere retrieval systems - can be enriched to interactive multimedia DL-systems and are combined into one virtual personal digital library.

OMNIS/2 is part of the Global Inventory Project of the G7 countries.

1 Motivation

The information available to the individual increased considerably over the last decade, in quantity as well as in the number of media types. The different kinds of data required different storage and retrieval systems. This led to different digital library systems, which today are powerful tools, but are not transparently connected. It is e.g. not possible to create a user defined link in any document of a digital library system to another document in another digital library although the source document contains a reference (e.g. bibliographic reference or meaningful word). Additionally, it is not possible for users to “work” with the libraries (i.e. to annotate books, create own documents, link documents, etc.) as they are retrieval systems only. In this paper we present our OMNIS/2 system which

¹ OMNIS/2 is funded by DFG (German Research Foundation) within the research initiative “V3D2” (“Distributed Processing and Delivery of Digital Documents”)

enables a user to make use of various information sources, i.e. digital library systems, and which combines them into one virtual personal digital library.

This paper is organized as follows. We start with an overview on missing features in current digital library systems in section 2. Section 3 includes related works on these topics. In section 4 and 5 we present the architecture of OMNIS/2 and explain some of the main internal concepts which are solutions to the mentioned missing features. The paper ends with a summary on further activities (section 6) and a conclusion (section 7).

2 Current State and Demands

Current digital library systems support various ways of retrieving documents, ranging from mere catalog retrieval (OPAC systems) and full text search, to content based search in media libraries like video libraries or music libraries. The following features, however, are available only in parts or are missing at all in current digital library systems:

- The integration of full text retrieval library systems and *multimedia* database systems, which may be distributed or remote, into a progressive, interactive, multimedia digital library system, which offers the option of transparently including other systems (with often large collections of data) and even introducing cross references among them.
- There is a need for a generic management of metadata, not only for documents which are stored in the multimedia digital library system, but also for any document of additionally connected digital library systems. This is necessary in order to make use of modern filtering and retrieval techniques and also for the transparent linking and processing of documents beyond the boundaries of a single digital library system.
- Automatically generated links between documents inside of a digital library and among documents of different digital library systems. Every retrieved article from a digital library system contains further bibliographic references which are mostly stored in the same digital library system. To retrieve these references the user usually has to initiate another search query, possibly with further restrictions, until the correct result is returned. An automatic linking between documents is not supported today.
- In addition to automatically generated links users want to create and follow personal relevant links to other documents in the same digital library system or to other systems. This feature is often asked for by research groups who work with specialized library systems like VD17 (a digital library of all German printings of the 17th century, a former project) [4].
- Current library retrieval systems lack the possibility of adding personal annotations to documents, as this would require write permission for all users of the digital library system. Annotations are helpful, though, to explore content for oneself or for a certain user group or to simply discuss a topic. Discussions require annotations on annotations. If various lengths and various types of media annotations are possible, a user friendly authoring tool

and the integration of a multimedia database system for storage purposes are required. Annotations are supported in some newly developed media libraries and personal libraries, but not in catalog and retrieval systems.

- Almost all current digital library systems miss a personalization scheme which every user can feed with personal interests. The personalization can, on the one hand, be used with an existing pull-technology for an additional semantic filtering of information or, on the other hand, as a push-technology to inform a user about new relevant incoming titles. The personalization feature should be adaptive to recognize shifts in the users' interests, but it should also be a corporate tool to create recommendations.

3 Related Work

The above mentioned missing features show that a simple library or document system is no longer sufficient for a comprehensive scientific work. Therefore several systems are emerging which search in different sublibraries or collections (e.g. Alexandria Digital Library [7], Stanford Digital Library Project [18], NC-STRL [14], DeLiver [6], Karlsruhe Virtual Catalog [11], Elektra [12], MeDoc [5], etc.). The *Karlsruhe Virtual Catalog* [11] and *Elektra* [12] e.g. query several underlying databases or external catalog and full text systems and merge the answers into one result, but they do not support multimedia documents, annotations, personal links or an adaptive personalization. The interoperability of digital libraries is an important topic in the *Stanford Digital Library Project*. Their developed InfoBus especially supports this by providing a general infrastructure. The commercial product *Hyperwave* [13], which is a webserver with database functionalities, offers techniques to index, to search, to link and to annotate documents which are stored *on* the server.

Today most frameworks and proposals prefer XML as a description and exchange language for meta data. Several companies extended their database systems recently to support the storage of XML documents. One of these systems is *Oracle8i*, which can be seen as a database system with integrated functionality of a webserver. The special properties of XML as semistructured data have led to the development of database systems which use new graph-oriented data models for semistructured data in contrast to the traditional relational or object-oriented approach. Among them are *Lore*, Stanford University [20] and *Tamino*, Software AG [15].

Metadata is a very important aspect if catalog data is maintained and if queries are exchanged among several systems. Standardized schemes for metadata are available in the field of libraries (e.g. MARC, MACHine REadable Cataloging, or the German MAB-Format). The *Dublin Core Metadata Initiative* deals especially with the definition of metadata for electronic resources. The recommendation of the Dublin Core Initiative, however, is difficult to use for multimedia applications (1:1 rule). A conceptual model for metadata architecture is available with the *Warwick Framework*. The Warwick Framework is not a technology or a technique but rather a set of principles for metadata. Parts of

the principles of the Warwick Framework are used in *RDF* (Resource Description Framework), which specifies how to use XML to represent metadata in the form of statements about properties and relationships of items on the Web.

4 Architectural Overview

4.1 Design Principles

Our idea is originating from the fact that we can not change or do not want to change the existing digital libraries as they are either well established systems or the effort to change them would be tremendous. We can consider the existing systems however as large containers of documents with a powerful query language. Therefore our extensions act as a metasytem for these digital library systems, enhancing them by new functionalities. Even if our metasytem will be integrated into one of the digital library systems in the future, this would be just an additional layer on top of the original system. Thus in an abstract view it would still be the same architecture. Our metasytem is able to search in one or various digital libraries and to automatically link their documents, to annotate them, to extend them with multimedia components and to personalize them. The original documents themselves remain in the original database systems and are never modified not even for added anchors. Documents are represented in the meta database simply by their logical addresses and meta data. The linking, including the anchor positions, is stored in the metasytem exclusively and is included dynamically into the retrieved documents at run-time. In the same way documents can be annotated with user-related, group-related or general annotations. These annotations can be arbitrary long texts, multimedia annotations or even annotations on annotations. In addition we have developed a concept for personalization (GRAS algorithm), which supports a better filtering and a personal ranking of the results of the underlying digital library systems. The GRAS algorithm unlike other personalization schemes uses Gaussian curves to describe user and object profiles, which results in a better overall quality of retrieval. The personalization feature can also be used as a messaging system about new relevant documents in the library. Both push and pull strategies are possible. The annotations and the data necessary for the personalization feature (i.e. user and object profiles to all documents) are also stored in the meta database. Often users like to create, to store and to retrieve additional personal documents together with the original documents. Since these personal documents may consist of texts, tables, graphics, audio or video the metasytem itself is required to offer all features of an interactive, multimedia database system including user workspaces, full text indexes, etc. For uniformity and consistency reasons the local documents are handled like documents in external digital libraries except for the write permission. If they are annotated or enriched with links, this is again stored outside the document in the meta database. This concept can be seen as an extension of the Amsterdam Hypermedia Model [9], which by itself is an extension of the Dexter Reference Model [8], clearly separating the document layer (within-component layer) from the linkage layer (storage layer). OMNIS/2

supports a more powerful link concept than the currently existing one in the WWW: n:m links, bidirectional links and newly developed temporal links and overlapping links are supported, while the consistency of links is still ensured.

Summarizing, the metasystem can, on the one hand, be seen as an integrated, interactive digital library system with full text, multimedia and hypermedia documents, and on the other hand, as a meta system for existing digital library systems, extending these by hypermedia elements including user workspaces, automatic and personal links, annotations, personal multimedia documents, a transparent access to further library systems, an integrated management of meta data and a unified interactive user interface.

4.2 Integration of External Systems

In general we can distinguish two possibilities of integrating an existing digital library system into OMNIS/2: A tight integration and an integration with the help of XML. The tight integration utilizes an internal interface to the relevant systems. Then the systems are directly integrated into OMNIS/2 by using these interfaces. If systems use the XML integration OMNIS/2 manages the shared DTD (document type definition) for XML and additionally uses the tagged information for its management of meta data. Integrity rules which are necessary for the storage and the processing of XML documents and DTDs in database systems are therefore an important part of OMNIS/2.

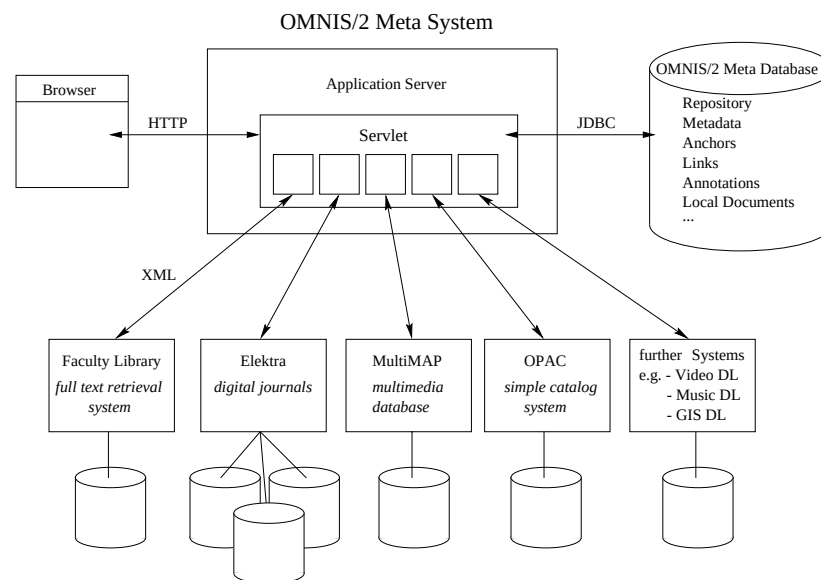


Fig. 1. Architecture of OMNIS/2

5 Implementation

5.1 Implementation Aspects of the Architecture

An important aspect of OMNIS/2 is that neither the integrated digital library systems nor the stored documents need to be modified. It is not even required to update the original documents in order to mark the links, respectively the anchors or the starting points of annotations. This is possible with the help of a three layer architecture similar to the Amsterdam Hypermedia Reference Model [9]. The integrated digital library systems correspond to the within-component-layer, which holds the documents (Fig. 1 shows the overall architecture). The composition, linking and annotation is performed in the meta layer, which consists of an application server and a database connection to the meta database. This layer represents the storage-layer of the Dexter and the Amsterdam Reference Model. The third layer is the presentation layer running in the user's web browser. This is the conceptional view on the architecture, which is shown in Fig. 1 vertically.

The integrated systems shown in Fig. 1 are only examples. In a more technical view on the architecture, which is shown in Fig. 1 horizontally in the upper part of the figure, the so called "three-tier architecture" becomes visible which is the foundation of today's implementations of the web access to the database. In the implementation we have chosen a server-side access to the meta database by using Java servlets and JDBC to lower the load on the clients's side (to support thin-clients). The servlet extends the webserver to a complete application server which holds the core of the metasystem. The communication of the browser and the webserver is performed via the standard HTTP port or by exchanging java classes, if applets are necessary for the presentation. User input is forwarded by the browser to the servlet and is processed there. All servlets are connected to the meta database with JDBC.

If a query is initiated for one of the coupled digital libraries then the query has to be transformed by the servlet into the URL-format of the relevant library system and sent to it. If a link is followed into a document of one of the coupled databases a corresponding query has to be generated by the servlet with the help of the information in the meta database. In the case of the XML coupling, the metasystem parses the retrieved XML documents with an XML parser and enriches them with the additional information from the meta database (like link anchors, annotations, etc.). To do this, a second query to the meta database is initiated. This query is always necessary. The resulting, enhanced document is currently transformed into HTML containing Java applets or plugin applications and is sent to the WWW browser for presentation.

In this context the architecture of the meta database plays a key role. It has to store metadata, link anchors, annotations of various media types and local documents which can also consist of different media types. The amount of stored data is expected to grow rapidly as anchor and link information has to be stored for every single object that was edited by users. Additionally the personalization feature requires the storage of object profiles for all documents in the meta

database. Except for the anchor values the data is strongly structured. It is important to note though that there is also semi-structured metadata from the underlying digital libraries if these systems use metadata in XML. The aspects which have to be taken care of in the meta database for semistructured data include storage, retrieval, internal representation, fragmentation and indexing. As the metasytem is based on several underlying systems and also maintains its own meta database inter-system consistency of the stored metadata and the underlying digital libraries is a very important aspect.

5.2 Proof of Concept

After the specification of the architecture we developed a demonstrator in order to validate the practical usability of our concepts. We integrated the digital library system of the Faculty for Computer Science at the Munich University of Technology with 80 000 full text documents as an established application to test the external integration via XML. In order to be able to access the retrieved data of the digital library system it had to be extended with an XML output right below its presentation layer. This was the only necessary extension to the digital library system, a surprisingly easy task and implemented within a short time. It is now possible to access the digital library via the OMNIS/2 layer. This results in a personal working environment (with links, annotations, etc.) and enhances the mere retrieval system.

5.3 Meta Database Schema and Efficiency

The meta database of OMNIS/2 is a very important part of the system. It not only contains the anchors and links of the documents, but also the annotations, the personal interests of the users and the metadata for the documents (local and external). As a result the development of the relational scheme of the database was a critical step in the development of OMNIS/2. In addition the database scheme, including the used index structure, decides upon the possible access methods and their efficiency. Especially the queries that deal with link navigation and full text search have to be processed efficiently. Our proposed database scheme can be subdivided into different areas. The areas cover (a) the components, (b) the linking, (c) the full text search and (d) the user administration.

A component in this context is a storable hypermedia object. The modelling of the links allows $n:m$ links. Due to efficiency they are, however, stored in $n*m$ 1:1 links and are referenced by a link_UID. The full text search covers full texts, all names and contents of all documents with a very efficient index-structure, supporting arbitrary truncations [1]. The user administration supports a user-based *channel* concept following the Amsterdam Reference Model. A number of different attributes for the presentation can be set globally and locally, thereby enabling a presentation for homogeneous media types.

During the development of the demonstrator the implementation of an efficient link navigation was an important point of research. Links are entities and thereby fulfill referential integrity, if they are not external links. The following

types of links are supported: n:m links, bi-directional links, temporal links and overlapping links. The Dexter and the Amsterdam Model are reference models in the sense of postulates. We implemented these models on top of relational and object-oriented database systems and examined the complexity for link navigation and a series of optimizations. Altogether the overall complexity of link navigation could be lowered from 3 square addends and 5 linear addends to $O(\log(n) + n \cdot \log(m))$ whereby $n = |\text{links}|$ and $m = |\text{anchors}|$. Thus our result is an efficient operation supporting the full Amsterdam functionality.

5.4 The Personalization and Filtering Feature

The goal of the personalization is to filter data, which is particularly relevant for the user from a large information base or to sort it according to personal importance or relevance. We have developed the Gaussian Rating Adaption Scheme (from now on referred to as GRAS) which is especially suitable for hypermedia documents. GRAS utilizes user and object profiles. Each profile consists of *several* topic profiles. Each topic profile describes the interest of a user or the contents of an object in a special range of topics. Simple rating schemes use only one evaluation value to describe the interest of a user in a topic. This, however, is not adequate in many cases. GRAS is a technique which uses Gaussian curves for the representation of the topic profiles. A Gaussian curve is characterized by two parameters μ and σ . One value describes the focus or average value (μ) of the interest of user on a scale from “is highly interested in this topic” to “does not like this topic at all”. The second value expresses the width (σ) of the interest of user. Figure 2a shows two Gaussian curves with different μ and σ values. μ shifts the Gaussian curve on the x-axis. σ creates a slimmer (lower σ value) or broader curve.

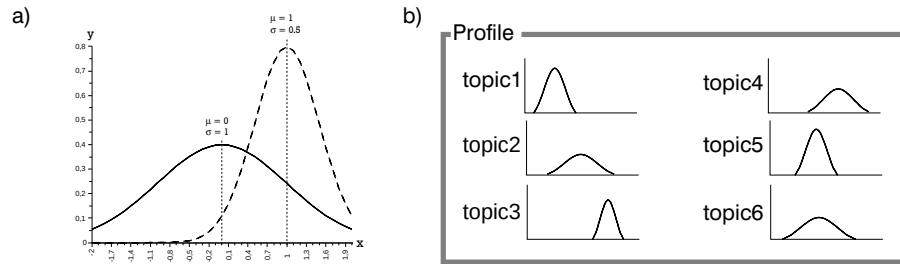


Fig. 2. a) example for two Gaussian curves; b) example of a profile consisting of several topic profiles.

The dotted curve can be interpreted as the interest profile of a very interested person. A highly positive μ value, combined with a low σ value, represents a high significance and a great interest in objects whose contents meet exactly the special topic. The other curve represents the interest profile of a user, who is

neither very interested in a topic nor completely averse. Therefore the μ value is 0 and the σ value is 1, which describes no special significance. Figure 2b shows, how several topic profiles represented by Gaussian curves form a user or an object profile.

The intersection of the two Gaussian curves (one for the object profile and one for the user profile) describes the significance of a user's interest in a certain object for a topic. Detailed analyses show that also negated interestes are modelled and treated correctly. The overlapping of the profiles leads to an n-tuple of overlapping values for topics for each object. Different decision models can be used to create an overall ranking: the *additive model*, the *conjunctive model*, the *disjunctive model* or the *average model* which we suggested.

The GRAS method was extended by an algorithm for the adaptive generation and adjustment of profiles, so that the user and object profiles correctly adjust themselves with feedback as time passes. For very large quantities of data the objects can therefore also be initialized with the standard Gaussian curve (which represents no special interest). This is a big advantage over methods which require an explicit assignment of topics for each document.

6 Further Activities

With the large sets of data available in current digital library systems it is no longer possible to add all links by hand with an authoring tool. An automatic creation of links eliminates this problem. The available options range from keywords, meaningful words or phrases which are stored in the meta database of OMNIS/2 and are matched against words in retrieved documents to heuristics which extract the necessary information from untagged literature references in scanned articles. We consider the further development and improvement of these techniques as a very promising topic for further research. With the development of the GRAS algorithm a powerful tool is available to implement the personalization feature. The integration of this tool is planned for the next version. It will then be combined with a planned notification service which enables users to be notified about new documents in the various digital libraries according to their personal interests or if annotations are added to their own annotations. This offers a push technology in addition to the existing pull technology.

7 Conclusion

We presented the concept and the architecture of the OMNIS/2 system. OMNIS/2 is a meta system for various existing digital library systems and equips users with a tool that enables them to search several systems at once and also to benefit from links between documents of different digital library systems. The links are generated automatically whenever possible but users are also able to add links by themselves. In addition it is possible to annotate external documents without a write permission for the library systems which hold the annotated

documents. The annotations are not limited to text and can consist of multimedia documents. Users will be able to use the personalization feature to create their own view on the documents and to “work” with digital library systems by themselves. In its current version the system has already access to 80000 documents of the digital library system of the Faculty for Computer Science at TU München. This means a clear shift from a mere retrieval system to a personal working environment.

References

1. Bayer R., *The digital library system OMNIS/Myriad*, Proc. 18th Australasian Computer Science Conference (ACSC'95), Glenelg, South Australia, Feb. 1995.
2. Baldonado M., Chang C. K., Gravano L., Paepcke A., *The Stanford Digital Library Metadata Architecture*, Int. Journal on Digital Libraries, 1(2), 1997, pp. 108-121.
3. Daniel, R., Lagoze, C., Payette, S. D., *A Metadata Architecture for Digital Libraries*, Proc. of ADL'98, Santa Barbara, CA, IEEE Computer Society, 1998, pp. 276-288.
4. Dörr M., Haddouti H., Wiesener S., *The German National Bibliography 1601-1700: Digital Images in a Cooperative Cataloging Project*, Proc. of ADL'97, Washington DC, IEEE Computer Society, 1997, pp. 50-55.
5. Endres A., Fuhr N., *The MeDoc Digital Library Operates as a Network of Distributed Servers*, Comm. of the ACM, 41(4), 1998.
6. Federated Repositories of Scientific Literature, University of Illinois at Urbana-Champaign, <http://dli.grainger.uiuc.edu>
7. Frew J. et al., *The Alexandria Digital Library Architecture*, Proc. of ECDL' 98, Crete, Greece, Springer Verlag, Sept. 1998, pp. 61-73.
8. Halasz F., Schwartz M., *The Dexter Hypertext Reference Model*, Comm. of the ACM, 37(2), Feb. 1994, pp. 30-39.
9. Hardman L. et al., *The Amsterdam Hypermedia Model: Adding Time and Content to the Dexter Model*, Comm. of the ACM, 37(2), Feb. 1994, pp. 50-62.
10. Kahabka T., Korkea-aho M., Specht G., *GRAS: An Adaptive Personalization Scheme for Hypermedia Databases*, Proc. of the 2nd. Conference on Hypertext - Information Retrieval - Multimedia (HIM '97), pp. 279-292.
11. Karlsruhe Virtual Catalog, Karlsruhe University, <http://www.ubka.uni-karlsruhe.de/hylib/en/kvk.htm>
12. Kowarschick W., Vogel P., Bayer R., *Elektra: An Electronic Article Delivery Service*, Proc. of the 8th Int. Workshop on Database and Expert System Applications (DEXA 97), Toulouse, France, IEEE Press, Sept. 1997, pp. 272-277.
13. Maurer H., *Hyper-G now Hyperwave*, Addison-Wesley, 1996.
14. Networked Computer Science Technical Reports Library (NCSTRL), Cornell University, <http://www.ncstrl.org>
15. Software AG, *Tamino White Paper*, 1999, <http://www.softwareag.com/tamino/>
16. Oracle Corporation, *XML Support in Oracle8i and Beyond*, Technical White Paper 1998, <http://www.oracle.com/xml/documents/xml.twp>
17. Paepcke A., Chang C. K., Garcia-Molina H., Winograd T., *Interoperability for Digital Libraries Worldwide*, Comm. of the ACM, 41(4), April 1998, pp. 33-43.
18. Stanford Digital Library Project: <http://diglib.stanford.edu/>
19. Specht G., *Complexity Analysis of Link Navigation in Dexter Based Hypermedia Database Systems*, Informatica, 8(1), Vilnius, Lithuania 1997, pp. 23-42.
20. Widom J., *Data Management for XML*, Stanford University, 1999, <http://www-db.stanford.edu/lore>

Modeling Archival Repositories for Digital Libraries^{*}

Arturo Crespo and Hector Garcia-Molina
{crespo,hector}@db.stanford.edu

Stanford University

Abstract. This paper studies the *archival problem*: how a digital library can preserve electronic documents over long periods of time. We analyze how an archival repository can fail and we present different strategies that help solve the problem. We introduce *ArchSim*, a simulation tool that for evaluating an implementation of an archival repository system and compare options such as different disk reliabilities, error detection and correction algorithms, preventive maintenance, etc. We use ArchSim to analyze a case study of an Archival Repository for technical reports.

1 Introduction

A continual threat to digital libraries is the loss of documents. Digital information can be lost not just through magnetic decay in storage devices, but also because of format and device obsolescence. This problem will only get worse as more and more information is provided only in digital form. The solution is an *Archival Repository* (AR), a system capable of storing and preserving digital objects (e.g., movies, technical reports) as technologies and organizations evolve [3]. An AR must preserve not only the bits, but also the “meaning” of its documents [9]. For instance, to preserve a “postscript” technical report, the AR needs to maintain the postscript bits, metadata indicating that this document is postscript, a program that can interpret and render postscript, and an environment that can execute the rendering program.

One of the main challenges in designing an archival repository is how to configure the repository to achieve some target “preservation guarantee” while minimizing the cost and effort involved in running the repository. For example, the AR designer may have to decide how many sites to use, what types of disks or tape units to use, what and how many formats to store documents in, how frequently to check existing documents for errors, what strategy to use for error recovery, how often to migrate documents to a more modern format, and so on. Each AR configuration leads to different levels of assurance, e.g., on the average a document will not be lost for 1000 years, or in 1000 years we expect to still have access to 99% of our documents. Each configuration has an associated cost,

^{*} This material is based upon work supported by the National Science Foundation under Award 9811992.

e.g., disk hardware involved, computer cycles used to check for errors, or staff running each site.

The number of options and choices is daunting, and the AR designer has few good tools to help. The traditional fault tolerance models and techniques, of the type used to evaluate hardware, are a helpful starting point, but they do not capture the unique complexities of ARs. For example, traditional models may have difficulty capturing different document loss scenarios (e.g., missing interpreter, missing bits, missing metadata) and they frequently assume failure distributions (e.g., exponential) that are too simplistic.

In this paper we present a powerful modeling and simulation tool, *ArchSim*, for helping in AR design. ArchSim can model important details, such as multiple formats, preventive maintenance, and realistic failure distribution functions. ArchSim is capable of evaluating a large number of components over very long time periods. ArchSim uses specialized techniques in order to run comprehensive simulations in a time frame that allows the exploration of different policies.

Of course, no model can be absolutely complete: there is an intrinsic trade-off between how detailed the model is and the complexity (even feasibility) of its study. In our case, we have chosen to ignore (at least in this initial study) information loss due to format conversion (migration to a new format is always successful and does not introduce any loss). We also do not model partial failures (a failure where we can salvage part, but not all, of the information in a device). As we will see, we also rely on an “expert” that can provide failure distributions for the components of the systems. For instance, if format obsolescence is an issue, an expert needs to give us the probabilities of different format lifetimes.

In summary, our contributions are:

- A comprehensive model for an AR, including options for the most common recovery and preventive maintenance techniques (Sections 2, 3 and 4).
- A powerful simulation tool, ArchSim, for evaluating ARs and for studying available archival strategies (Sect. 5).
- A detailed case study for a hypothetical Technical Report repository shared between two universities. Through this case study, we evaluate AR factors such as disk reliability, handling of format failures, and component aging.

2 Archival Problems and Solutions

As a first step in modeling and evaluating archival repositories (ARs), it is important to understand how information can be lost, and what techniques can reduce the likelihood of loss. Because of space limitations, we limit ourselves to a brief review of failure sources and avoidance techniques. The expanded version of this paper [8] contains the full review.

2.1 Sources of Failures

An AR fails to meet its guarantee when it loses information. Such a loss may be caused by a variety of undesired events, such as the failure of a disk, on an

operator error. A document is lost if the bits that represent it are lost, and also if the necessary components that give meaning to those bits are lost. We define the *components* of a document to be all the resources that are needed to support access to the document, e.g., the document bits, the disk that stores the bits, and the viewer that interprets the bits.

An undesired event does not necessarily cause information loss. For instance, if the AR keeps two copies of a document, and the disk holding one of the copies fails, then the document is not lost. It would take a second undesired event affecting the second copy to cause information loss. A document is *damaged* if a copy or instance of one of its components is corrupted or lost. Damaged documents should be *repaired* by the system to protect them from further failures.

The most common undesired events that may lead to the loss of a document can be broken down into the following categories. (We do not consider transient failures, e.g., a power failure, that do not lead to a permanent loss.)

Media decay and failure: Example: disk magnetic decay; a worn out tape breaks when it is being read.

Component Obsolescence: Example: we do not have a machine that can read a tape, even if the media is still readable.

Human and Software Errors: Example: a person or a program deletes a document; a program improperly modifies the files and data structures that represent the document. (Reference [11] suggests that humans and software are the most serious sources of failures in computer systems.)

External Events: Examples: fires, earthquakes and wars. Such events may cause several AR components to fail simultaneously. For example, a flood can destroy a collection of disks at one site. If a document was replicated using those disks, all copies will be lost.

2.2 Component Failure Avoidance Techniques

There are two well-known ways to avoid an AR failure: We can either reduce the probability of component faults, or we can design the AR so those faults do not result in document loss. In this subsection, we review techniques in the first category, while the next subsection will cover the second category. We organize the presentation in this subsection using the taxonomy of faults presented in Sect. 2.1.

Media decay and failure: To reduce the likelihood of media failure, we can store our document on very reliable media (e.g., use CDs instead of tapes). We can also ensure that the media is maintained in the best conditions possible. For example, by placing tapes in a low-temperature/low-humidity environment, we may increase their life by an order of magnitude [2].

Component Obsolescence: Reducing component obsolescence is hard, as it requires an accurate prediction of what operating systems, document formats, and devices will be used in the future. Still, use of standards, self-contained media (i.e., media that includes its own reader) [1], and equipment preservation may help.

Human/Software Errors: Good coding techniques can reduce the likelihood of these failures, e.g. by defining interfaces that minimize the amount of damage that can be done. In addition, program validation and operator training can help.

External Events: Damage from external or environmental events can be reduced by fire-proof walls, earthquake resistant buildings, and so on.

2.3 System-Level Techniques

Migration: Migration involves replacing a document component by a new, safer one. For example, suppose that “Postscript” readers are becoming obsolete and are being replaced by new “Postscript II” readers. Then, we may decide to migrate Postscript documents into the new format before Postscript readers become unavailable. Migration is particularly effective for storage devices. However, migrating to new formats is more challenging because some information may be lost in the transformation.

Replication: Replication makes copies or creates new instances of needed components, to ensure the long term survival of the component. Replication is used by many archival system including the Computing Research Repository [12], the Archival Intermemory Project [10, 5], and the Stanford Archival Vault [6].

Emulation: Emulation involves re-creating all the components needed to access a document on a new platform [16]. Emulation can be done at several levels: hardware, operating system, or application software.

Fast Failure Detection and Repair: With most system fault tolerance techniques, we need to check periodically for failed documents. Fast failure detection and repair yields improved reliability. For example, if one of two component copies have failed, the sooner we detect the problem and generate a new copy, the more protection we get against a second fault.

3 Architecture of an Archival Repository

Our goal in this section is to identify the elements of a typical archival repository (AR), so we can model each element and study how it impacts reliability. A typical AR stores documents in a *data store* that can fail. The AR can still achieve long-term survivability by enhancing the data store with an *archival system* (AS) that implements some of the techniques of Sect. 2.3. We present our AR model in Fig. 1. The figure shows the AS modules (in solid-line boxes), the non-fault-tolerant store (in a dashed-line box), and the archival documents. The arrows represent the runtime interactions between the elements.

3.1 Archival Documents

In our architecture, an archival document embodies information. An archival document cannot be just a bag of bits, but it must also include all the components necessary to transform the bits into a human comprehensible form.

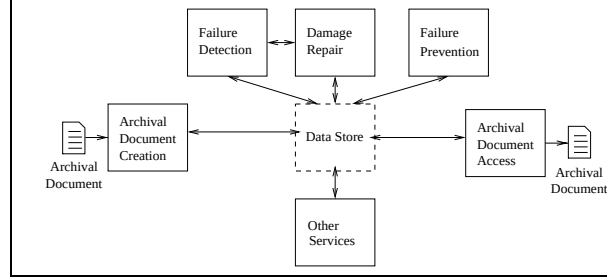


Fig. 1. Archival Repository Model

An archival document is an abstract entity. The connection between a document and access to it is achieved through *materializations*. A materialization is the set of all the components necessary to provide some sort of human access to a document. For example, a materialization may include the bits, disks, and format interpreters necessary to display a technical report. The same technical report may be accessible through a different materialization, that may include a different format interpreter to print the technical report. The AR is able to preserve documents by preserving the materializations and their components. In this paper, we treat the documents as “black boxes.” We do not attempt to take advantage of document structure (e.g., chapters in a book).

3.2 Architecture of the Non-Fault-Tolerance Store

The *Store* encompasses the set of components, such as sites, disks, or format interpreters that make materializations accessible. Because the store is not fault tolerant, materializations may be lost. A materialization is considered lost when *any* of its components has failed. If *all* of the materializations of a document are lost, then the document is considered lost.

To create a materialization, first we must ensure that the necessary components exist in the store. Then we create a record that links together the components as a materialization. For example, say we want to create a document materialization that requires the bits in file “doc.ps”, which are located in *disk*₂ in *site*₁ and requires the postscript interpreter. Any components that do not exist already (e.g., the file doc.ps) are created. In some cases, “creating” a component may require a physical action, e.g., adding a new printer or disk to the store. Once the components exist, a metadata record containing references to the components is created (e.g., $\langle site_1, doc.ps, disk_2, postscript \rangle$). This record serves as the identifier for the materialization.

A document metadata record includes the records for all available materializations. Note that a document is not accessible if its metadata record is corrupted or lost. Therefore, the record must be one of the required components for any materialization.

3.3 Architecture of the Archival System

The AS provides fault tolerance by managing multiple materializations for each document. The AS monitors these materializations, and when a failure is detected, attempts to repair them. The AS can improve fault tolerance further by taking preventive actions to avoid failures. The AS provides the user the ability to create and retrieve archival documents. It also provides miscellaneous services such as indexing, security, and document retirement, among others. When a user requests a document, the AS uses its metadata to find all the available materializations of that document, selects one and returns it to the user. In this section, we describe the six modules that make up the AS (see Fig. 1).

The *Archival Document Creation module* (ADC) generates new documents, implementing policies on the number and types of materializations that are needed. The main objective of this paper is to provide a framework that allows a system administrator to choose the best materialization policies to achieve a desired level of reliability.

The *Archival Document Access module* (ADA) services request for documents. Basically, the module translates the request for a document into a request for the appropriate components of one of the document materializations.

The *Failure detection module* (FD) scans the store looking for damaged or lost materializations. When a damaged materialization is found, the failure detection module informs the Damage Repair module (described below).

The *Damage Repair module* (DR) attempts to repair damaged documents. There are many strategies to repair a damaged document, as discussed in Sect. 2.3. The input of the DR module is a signal from the FD module.

The *Failure Prevention module* (FP) scans the store and takes preventive measures so materializations are less likely to be damaged. For example, the FP module may copy components that are stored on a disk that is close to the end of its expected life, into a newer disk.

Finally, the *Other Services module* (OS) provides miscellaneous services such as indexing, security, and document retirement. Retiring a document involves removing from the store any components that are no longer needed, even by other materializations.

4 Failure and Recovery Modeling

In this section, we will model the failure and recovery characteristics of an AR, based on the architecture presented in the previous section. First, we will explore how to model a non-fault-tolerance store, and then the archival system (AS). Later, we will combine these two models into an archival document model.

4.1 Modeling a Non-Fault-Tolerance Store

To model the failure characteristics of a store, we start with an abstract representation of materializations and components. We model a materialization as an

n-tuple $\langle mat_{id}, doc_{id}, comp_1, \dots, comp_n \rangle$; where mat_{id} is the materialization identifier, doc_{id} is the document identifier, and $comp_1 \dots comp_n$ are the components required to provide the required document access. The identifiers mat_{id} and doc_{id} together, form a unique id for the materialization. For example: $\langle M1, TR1233, doc.ps, site_1, disk_2, postscript \rangle$ means that the materialization $M1$ contains the document identified by $TR1233$ that needs the bits in file $doc.ps$, $disk_2$, $site_1$ and the postscript interpreter in order to be readable. A document can have more than one materialization. For example, Technical Report 1233 can also have the materialization $\langle M2, TR1233, doc.ps, site_1, disk_3, postscript \rangle$, which would be a copy of $M1$ but on a different disk ($disk_3$).

We model components by a tuple $\langle component_{id}, type \rangle$, where $component_{id}$ is a unique identifier for the component instance and $type$ is the class (e.g., file, disk, interpreter) to which the instance belongs.

To further model components, we need to describe:

- How many component instances and types are present in the system: this is, how many disks, formats, etc., are available.
- Failure distribution of each component type. Many components have two different failure distributions, one during archival and another during access. For example, a tape is more likely to fail when it is being manipulated and mounted on a reader than when it is stored. Therefore, each component will have two failure distributions: during archival (i.e., time to next failure when the component is not used) and during access. For some components, such as disks or sites, the access and archival distributions will be the same; but for other components, such as tapes or diskettes, these distributions can be very different.
- Time distribution for performing a component check. This distribution describes how long it takes to discover a failure (or to determine that a component is good), from the time the check process starts. For example, consider checking a tape. This may involve getting the tape from the shelf, mounting the tape, and scanning the tape for errors.
- Time distribution for repairing a component failure. This distribution describes how long it takes to repair a component. This distribution may be deterministic (if the component can be repaired in a fixed amount of time). Repair time may be “infinite” if the component cannot be fixed.

In addition, there is an important interdependency between components. Specifically, the failure of one component may cause the failure of another component. For example, if a site fails (e.g., because it was destroyed by a fire), then all the disks at the site will also fail. This failure dependency is captured by a directed graph. For example, an arrow between “Site A” and “Disk 1” in the interdependency graph means that if “Site A” fails, then “Disk 1” will also fail.

We close this subsection with two comments. First, we do not claim that the model presented for the store is complete. For instance, we have not included policies for handling concurrent access. There is always a tradeoff between complexity of the model and our ability to analyze it. We believe that our model

strikes a good balance in this respect, and captures the essential features of a store. Second, the reliability predictions we make are only valid for the *current* configuration of the repository. Over time, the repository will change (e.g., as new devices are introduced), so we may need to change our repository model. As the model changes, we may need to revisit our predictions.

4.2 Modeling an Archival System

In this section, we describe how to model the behavior of the modules of the archival system. We do not include failure distributions for these modules as we are assuming that the AS itself does not fail. We recognize that this is a strong assumption, but in this paper, we have chosen to concentrate on the failure of components, instead of on the failure of the system that provides fault-tolerance. How to develop error-tolerant robust software design have been study in [15].

Because of space constraints, we cannot describe each module separately. In general, we model the input of the modules with probability distribution functions and their behavior by algorithms. For example, consider the document creation module. Its input distributions tell us how frequently requests for document creation arrive, how many materializations each new document will have, and which components will be selected to participate in a given materialization. The algorithms for the failure detection module spell out what policies are implemented, e.g., if all components are checked on a regular basis or not.

The probability distributions that drive the model can be obtained in different ways. If we have data from a real system, we can use the data directly (trace driven), or we can define an empirical distribution, or we can fit the data onto a theoretical distribution [13]. If we do not have real data, we need to choose a theoretical distribution that matches our intuition. A sensible distribution to choose (when requests are generated independently) is a Poisson distribution for event inter-arrival times [4].

5 ArchSim: A Simulation Tool for Archival Repositories

To evaluate a possible AR configuration, we need to predict how well it protects documents. This prediction can sometimes be done analytically, but as the AR gets more complex, an analytical solution is impractical (and sometimes impossible). Instead, we rely on a specialized simulation engine for archival repositories: ArchSim. We start this section by discussing the specific challenges confronted when simulating an AR. Then we describe ArchSim and its libraries.

5.1 Challenges in Simulating an Archival Repository

ArchSim builds upon existing simulation techniques for fault-tolerant systems. However, the unique characteristics of ARs make their simulation challenging:

- *Time Span*: The life of an archival system is measured in hundreds, perhaps thousands of years. This means that simulation runs will be extremely long, so special precautions must be taken to make the simulation very efficient. Furthermore, given these long periods, failure distributions must take into account component “wear-out.” (A component is more likely to fail after 50 years than it is when new.) Simple failure distributions (e.g., exponentially distributed time between failures) are frequently used in fault-tolerant studies, but they cannot be used here since they do not capture wear out.
- *Repairs*: In an archival system we cannot in general assume that damaged components can always be replaced by new identical components (another common assumption when studying fault-tolerant systems). For example, after say 100 years, it may be impossible or undesirable to replace a disk with one having the same failure characteristics.
- *Component models*: Component models are fairly rich, compounding the number of states that must be considered. For instance, as we have discussed, a file is not simply correct or corrupted. Instead, it can be corrupted but the error undetected, it can be correct but not accessible for reads, and so on. The failure models in each of these states may be different, e.g., a file is more likely to be lost when being read.
- *Sources of failures*: A document can be lost for many reasons, e.g., a disk fails or a format becomes obsolete. Each of these failures has very different models and probability distributions. The approach of finding the “weak link” and assuming that all other factors can be ignored is not appropriate for ARs.
- *Number of Components*: An AR needs to deal with a large number of components and materializations. The challenge of simulating large number of objects has been studied extensively [14] and ArchSim uses those results.

5.2 The Simulation Engine: ArchSim

ArchSim receives as input an AR model, a stop condition (stop when the first document is lost or when all documents are lost) and a simulation time unit (minutes, hours, days, etc.). ArchSim outputs the mean time to failure (mean time to stop condition), plus a confidence interval for this time. We are currently considering other output metrics, e.g., the fraction of the documents that are available after some fixed amount of time. However, these other metrics are not used in our case study (Sect. 6).

For defining the AR model, each distribution and policy is implemented as a Java object, so they can be as general as necessary. For example, for component repair, the corresponding Java module can simply use a probability distribution (perhaps one of the library functions described below) to generate the expected repair time. However, that module can easily be replaced by one that first decides if the component can be repaired (using one probability distribution), and then for each cases generates a completion time (when the component is repaired or the repair is declared unsuccessful). ArchSim makes available a library of pre-defined failure distributions, that can be used to describe AR components.

5.3 ArchSim's Implementation

ArchSim follows the structure of a traditional simulation tool. Each module of the AR model can register future events in a timeline. For example, when a disk is created, the simulation uses the disk failure distribution to compute when the disk will fail; then, it registers the future failure event in the timeline. The simulation engine advances time by calling the module that registered the first event. This module may change the state of the repository and register more events in the timeline. After the module returns, the simulation advances to the next event, in chronological order. The user can choose between two end conditions for the simulation: the simulation can stop after the first document is lost or after all the documents are lost.

ArchSim needs to be very flexible and efficient to meet the challenges of simulating an AR. Flexibility is needed to model very different archival conditions and implementations. Speed is needed to cope with many materializations, components, and events. Additionally, each simulation needs to be run many times in order to obtain narrow confidence intervals.

In an AR simulation, many events are inconsequential. For example, suppose that the detection module schedules periodic detection events. If the detection event finds a fault (i.e., there was a failure event before the detection event), then the module starts a component repair; if no failure is detected, then the module does nothing. If a repair module checks a component with mean time to failure (*MTTF*) of 20 years, every 15 days, then, in average, 486 events ($20 * 365/15$) will be fired and the repair module will just return without doing anything; only in the event 487, when a failure of the component has happened, will the repair module perform an action. Given the large number of components and modules that may be part of the model, this large number of inconsequential events represents a significant overhead. To avoid this overhead and to improve efficiency, modules are allowed to register *conditional* events in the timeline. These events will only happen if some other event happened before them. By using conditional events, we can condition the firing of the detection event only if a failure event on a specific component happened before it. A conditional event is not registered directly in the timeline. Instead, it is registered in an index that is part of its triggering event. For example, if event B is conditional to the occurrence of event A, we will put B in the index of A. When a new event A is scheduled in the timeline, we look in the index and find that B is conditional to it, so at that point we also schedule B.

To reduce the number of events further, ArchSim also modifies failure distributions. For example, when modeling preventive maintenance, a large number of “replace component” events are generated. We can eliminate all those events, by modifying the failure distribution. Specifically, the original failure distribution is used to generate the time of the next failure of the component. If this time is higher than the prescribed preventive maintenance period, we ignore this time, and we generate a new failure time again using the original distribution. We repeat this process until a failure time is lower than the preventive maintenance

period. The new distribution then returns the number of iterations minus one, times the PM period, plus the last failure time.

Another challenge for ArchSim is how to deal with a large number of materializations and components. This is done by scheduling only the *next* failure for each component type and associating a trigger with that event. When the failure event happens, the trigger is activated. The trigger computes when the *next* failure of a member of that component type will occur, and adds it to the timeline. Obviously, this approach is only beneficial if we have many components of the same type, which is a reasonable scenario for an AR. In the case when we have N different distributions for N components, this technique does not improve the simulation time, but it does not increase it.

6 Case Study: MIT/Stanford TR Repository

In this section, we use ArchSim to answer some design questions for an hypothetical MIT/Stanford Technical Report Archival Repository. The AR follows loosely the Stanford Archival Vault (SAV) design [7] and implementation [6]. (We actually considered creating such a repository some years ago, when both institutions were participating in the DARPA sponsored CSTR Project.) In this case study, MIT and Stanford preserve their Computer Science Technical reports by replicating the reports at both universities. In this case study we will have to make many assumptions. Our goal here is *not* make any specific predictions, but rather to illustrate the types of evaluations that ArchSim can support, the types of decisions that must be made to model an AR, and the types of comparisons than can be made to support rational decisions among alternatives.

We will assume that the collection has 200,000 documents and that each document is stored in one or more of four available formats. The repository will have two types of components: storage devices (disks) and format interpreters. To ensure preservation, the AR maintains four materializations of each technical report; two materializations at Stanford and two at MIT. Materializations are created by choosing two formats out of the four available formats. Two of the materializations will be in one of the chosen format, while the other two will be in the other. Then, at each site, we place two materializations that are in different formats in two different disks. Figure 2 illustrates the arrangement of the technical reports in this system.

Disks, formats, and sites have uniform failure distributions with parameters $1/\phi_{sto}$, $1/\phi_{form}$, $1/\phi_{site}$. As our base values, we are assuming ϕ_{sto} , ϕ_{form} , and ϕ_{site} , the mean time to failure (MTTF) for disks, formats and sites, to be 3, 20, and 45 years respectively.

These values are our best guess for a typical archival system. We chose 3 years as the disk MTTF, as this is the normal period under which a hard drive is under manufacturer guarantee. We chose 20 years for format MTTF as we theorize that it will take that long for a well-known format to be replaced by a new format and for all displays and transformers for the old format to be lost. We chose 45 years for the site MTTF as we assign a 50% probability to the

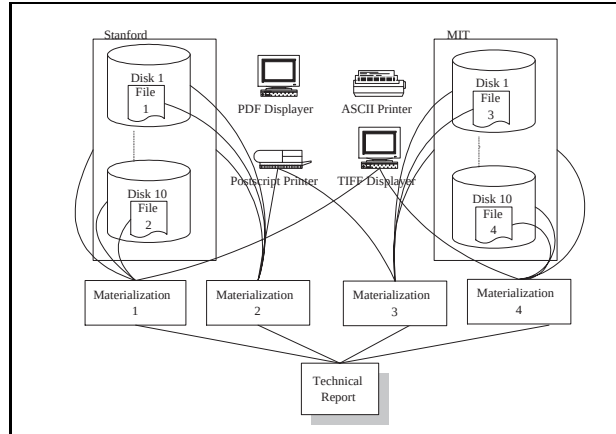


Fig. 2. MIT/Stanford CSTR Scenario

event of losing the Stanford site due to a high intensity earthquake in the San Francisco Bay Area (which is predicted to happen in a 45 year period).

The archival system checks for faults periodically. When a fault is detected in one of the storage devices, the bad device is retired, and a new device is set online. Then, the system regenerates the bad device by making a copy of the lost materializations from the other site. Similarly, when a format becomes obsolete, a new format is selected and a new set of materializations (transformed from a non-obsolete format) is created in the new format. In addition, in case of site failure, the site is recreated from the other site. If all sites, formats and devices that support all the materializations of a technical report are lost, then the technical report is lost and the simulation stops. Disks, formats, and sites are checked and repaired (if needed) every ρ_{sto} , ρ_{form} , and ρ_{site} days. As our base values, we are assuming ρ_{sto} , ρ_{form} , and ρ_{site} to be 60, 60, and 7 days.

To justify these values, we need to describe what is involved in the detection and repair of component failures. Detecting a failure in a disk involves scanning the whole disk and checking for lost data. When we find lost data, we need to order a new disk and then copy all the data that was in the damaged disk from other sources into the new disk. Assuming that the repair time is 60 days means that we need to dedicate only 3% of the disk bandwidth to scan all materializations in order to detect failures. We did not choose a quicker repair because the scanning overhead would be too high in our opinion. For example, 25% of the disk bandwidth is required to detect failures in 7 days.

For formats, the detection/repair times imply that we are able to realize that a format is obsolete and that we can create a new copy of the document from a non-obsolete format within a 60 day period. In the case of site failures, we are assuming that the detection/repair time for the site is much lower than for formats and disk. The detection itself should be rather fast in this case (the entire site is down), and the 7 days could be the time it takes to find a backup site to take over.

<i>Parameter</i>	<i>Symbol</i>	<i>value</i>
Number of disks	n_{sto}	100 per site
Number of formats	n_{form}	4
Number of documents	num_{doc}	200,000
Mean Time to Disk Failure	ϕ_{sto}	3 years
Mean Time to Format Failure	ϕ_{form}	20 years
Mean Time to Site Failure	ϕ_{site}	45 years
Disk Failure Detection/Repair time	ρ_{sto}	60 days
Format Failure Detection/Repair time	ρ_{form}	60 days
Site Failure Detection/Repair time	ρ_{site}	7 days

Fig. 3. Base values

In this case study, we assume that failures are total. This is, we cannot partially repair a component and salvage some of the materializations. The failure distribution during access will be assumed to be the same as the failure distribution during archival. The base values for our simulation are in Fig. 3. The simulation parameters are detailed in the the expanded version of this paper [8]. That extended paper also contains additional experimental results.

In our first experiment, we evaluate the effect of the failure MTTF and repair times of storage devices on the system MTTF. In Fig. 4, we show the system MTTF for different disk MTTFs, given a detection/repair time of 60 days. To single-out the influence of storage device failures, we are assuming in this experiment that formats and sites never fail. The dotted lines in the figure represent the 99% confidence interval for the simulation, while the solid line is the average of all the simulation runs. As expected, the system MTTF increases when the disk MTTF increases (when the disk failure rate decreases). The exponential shape of the curve is the result of the constant repair time. As we keep increasing the disk MTTF, it is much more improbable that another device will also fail before 60 days have passed. This graph allows us to select a good storage device for a target system MTTF. If the library targets a 10-year MTTF, then a disk with a failure MTTF of 3 years will suffice. However, if the library requires a MTTF of 100 years, then we will need disks with a MTTF of about 6 years. Most manufacturers guarantee their disks for three years and very few guarantee them beyond five years. Therefore, a 6-year disk MTTF requirement will be hard (or very expensive) to meet. Nevertheless, we can still achieve our target system MTTF by changing other parameters in our system, as we will see in the next experiment.

We now evaluate the sensitivity of the system MTTF to the repair time (ρ_{sto}). In Fig. 5, we show the MTTF of the AR for different disk MTTF (ϕ_{sto}) values, and for different expected detection/repair times. (Note that the graph has a logarithmic scale. Confidence intervals are not shown to avoid clutter.) First, let us concentrate in the curve for a ϕ_{sto} of 3 years. As expected, the MTTF of the system decreases when the ρ_{sto} of the storage devices increases. However, the shape of the curve is more interesting. At low detection and repair times, there is a high positive impact on the MTTF of the system. However, as we increase the repair times, the system MTTF drops sharply. Interestingly, for

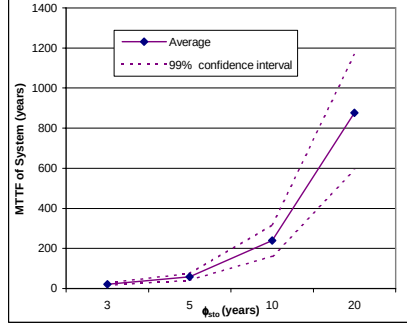
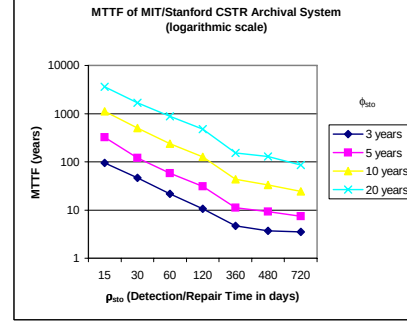
Fig. 4. MTTF vs. ϕ_{sto} with $\rho_{sto}=60$ days

Fig. 5. System MTTF (logarithmic scale)

a repair times greater than 120 days, about 1/9 of the MTTF of the storage device, the effect on the system MTTF of the detection/repair module is small. Thus, the detection and repair times must be much lower than the storage device failure MTTF to have a significant effect on the MTTF of the Archival System.

What is the optimal solution for a given MTTF with respect to disk MTTF repair times? The answer depends on the cost assigned to those two factors. By looking to all the curves of Fig. 5, we can observe that we can achieve similar MTTF by using better media or by reducing the detection/repair times. For instance, a system that uses a storage device with ϕ_{sto} of 5 years (i.e., a low quality storage device) and has a detection/repair time of 30 days, is as good as a system that uses a high quality storage device with ϕ_{sto} of 20 years, but is only checked and repaired every 360 days. The decision of which alternative to choose will depend on the cost of the storage device versus the cost of more frequent detections.

We now expand our experiments by allowing formats to fail. In Fig. 6 we show system MTTF as a function of ϕ_{sto} when formats can fail. We fix ρ_{sto} at 60 days, and now formats can fail with $\phi_{form} = 20$ years. To avoid introducing additional factors in the analysis, we assume site failures still cannot happen. Format failures are detected and repaired with $\rho_{form} = 60$ days. For comparison purposes, we have included in Fig. 6, the results presented in Fig. 4. The important conclusion that we can derive from the figure is that in an AR we cannot focus on single component types. It is surprising that even though the format MTTF is much larger than the disk's MTTF, the failure of formats still has a significant impact. This is because a document is lost if there is a disk failure *or* a format failure. The result is that we are taking the “worst” of those two failures, resulting in a system with a low MTTF. The result of this experiment shows that we need a comprehensive model, like the one proposed in this paper, to realize the interactions between the components and their effects on system MTTF.

In the next experiment, we use our system to explore a more complex failure distribution for storage devices that includes the issue of “infant mortality.” When the failure distribution includes infant mortality, storage devices have a higher failure rate in the beginning of their life (in our case, 30 days) than in

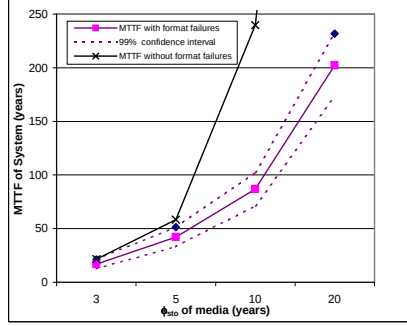


Fig. 6. MTTF with format failures

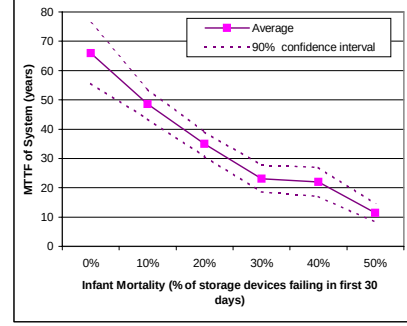


Fig. 7. MTTF with Infant Mortality

the rest of their lives. This can be expressed as a distribution that has a low MTTF within the first 30 days and a higher MTTF after that. The MTTF after the 30-day period will be 5 years. We will vary the early MTTF between 44 and 285 days. In this experiments we will assume that formats cannot fail. All other assumptions and repair procedures of the previous experiments are maintained (see Fig. 3). In Fig. 7 we show the system MTTF at given percentages of storage devices that fail in the first 30 days. For example, with an early MTTF of 285 days, 10% of the devices will fail within 30 days. With a MTTF of 135 days, 20% will fail. As expected, the higher the infant mortality, the lower the MTTF of the system. At a 0% infant mortality, the system MTTF was 65 years, dropping to 48 years when the infant mortality is 10% and dropping to only 11 years at 50% infant mortality level. Given this, when using components that suffer from infant mortality, the failure detection module should check new components much more often than older components.

We now explore the issue of aging of storage devices. With aging, a storage devices will have a lower MTTF at the end of its life. For example, a disk may have a 5 year MTTF during its initial life and a MTTF of 2 years when it reaches its “aging” phase. In this experiment we will assume that formats cannot fail. All other assumptions and repair procedures of the previous experiments are maintained (see Fig. 3). We evaluated the MTTF of the system for different points when aging starts (no graph shown). As expected, the sooner aging starts, the lower the MTTF of the system. If aging never occurs, the MTTF of the system is 65 years. If aging starts after 1 year, the system MTTF is 17 years, increasing to 42 years when aging starts after 5 years. Given this, when using components that suffer from aging, a way to increase the MTTF of the system is for the failure detection module to check old components much more often than newer components. Moreover, we should consider replacing old components with newer ones before the old components fail.

7 Conclusions

In this paper, we have studied the *archival problem*. We studied the different options for recovery and preventive maintenance, developing a comprehensive

model for an AR. We described a powerful simulation tool, ArchSim, for evaluating ARs and the available archival strategies. We described how ArchSim can efficiently perform large simulations many components and very long durations. We demonstrated the use of ArchSim with a case study for a hypothetical Technical Report repository operated between Stanford and MIT. We considered options such as disks with different reliability, number of copies, format failure handling, and preventive maintenance. We believe ArchSim can help librarians and computer scientists make rational decisions about preservation, and help achieve better archival repositories.

References

1. Douglas A.Kranch. Preserving electronic documents. In *ACM Digital Library Conference*, 1998.
2. J.W.C. Van Bogart. *Magnetic Tape Storage and Handling*. Natl. Media Lab, 1995.
3. C. Borgman, S. Chen, H. Garcia-Monlina, K. Thibodeau, , and G. Wiederhold. *NSF Workshop on Data Archival and Information Preservation*. National Science Foundation, March 1999. At <http://cecssrv1.cecs.missouri.edu/NSFWorkshop/>.
4. E. Çinlar. *Introduction to Stochastic Processes*. Prentice-Hall, 1975.
5. Yuan Chen, Jan Edler, Andrew Goldberg, Allan Gottlieb, Sumeet Sobti, and Peter Yianilos. A prototype implementation of archival intermemory. In *Proceedings of the Fourth ACM International Conference on Digital Libraries*, 1999.
6. Brian Cooper, Arturo Crespo, and Hector Garcia-Molina. Implementing a reliable digital object archive. In *Proceedings of the Fourth European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, 2000.
7. Arturo Crespo and Hector Garcia-Molina. Archival storage for digital libraries. In *Proceedings of the Third ACM International Conference on Digital Libraries*, 1998.
8. Arturo Crespo and Hector Garcia-Molina. Modeling archival repositories for digital libraries. Technical report, Stanford University. At <http://www-db.stanford.edu/~crespo/papers/ArchSimFull.ps>, 1999.
9. J. Garrett and D. Waters. Preserving digital information: Report of the Task Force on Archiving of Digital Information, May 1996.
10. Andrew Goldberg and Peter Yianilos. Towards an archival intermemory. In *Advances in Digital Libraries*, 1998.
11. Jim Gray and Andreas Reuter. *Transaction Processing: Concepts and Techniques*. Morgan Kaufmann Publisher, 1993.
12. Joseph Halpern and Carl Lagoze. The Computing Research Repository: Promoting the rapid dissemination and archiving of computer science research. In *Proceedings of the Fourth ACM International Conference on Digital Libraries*, August 1999.
13. A.M. Law and W.D. Kelton. *Simulation Modeling & Analysis*. McGraw-Hill, 1991.
14. T.I. Oren. Application of system theoretic concepts to the simulation of large scale adaptive systems. In *Proceedings of the 6th Hawaii International Conference on Systems Sciences*, pages 435–437, 1973.
15. D. K. Pradhan. *Fault-Tolerant Computer System Design*. Prentice Hall PTR, 1995.
16. Jeff Rothenberg. Avoiding technological quicksand: Finding a viable technical foundation for digital preservation. Technical report, Council on Library and Information Resources (CLIR), Washington DC, 1999.

Implementation and Analysis of Several Keyframe-Based Browsing Interfaces to Digital Video

Hyowon Lee, Alan F. Smeaton, Catherine Berrut,
Noel Murphy, Seán Marlow, Noel E. O'Connor

Centre for Digital Video Processing
Dublin City University, Glasnevin, Dublin 9, Ireland
{hlee, asmeaton}@compapp.dcu.ie

Abstract. In this paper we present a variety of browsing interfaces for digital video information. The six interfaces are implemented on top of Físchlár, an operational recording, indexing, browsing and playback system for broadcast TV programmes. In developing the six browsing interfaces, we have been informed by the various dimensions which can be used to distinguish one interface from another. For this we include layeredness (the number of "layers" of abstraction which can be used in browsing a programme), the provision or omission of temporal information (varying from full timestamp information to nothing at all on time) and visualisation of spatial vs. temporal aspects of the video. After introducing and defining these dimensions we then locate some common browsing interfaces from the literature in this 3-dimensional "space" and then we locate our own six interfaces in this same space. We then present an outline of the interfaces and include some user feedback.

1 Introduction

The technologies used to capture, compress, store and transmit TV-quality digital video are now so far advanced that the development of digital libraries of video information is now an achievable target for many organisations. In our research group we are developing Físchlár¹, a system which currently captures, stores, indexes and allows browsing of broadcast TV material to users within our organisation [1]. Using a dedicated video server we can store almost 4 weeks of continuous content which we can stream to over a hundred users, concurrently. Our work is currently based around indexing and browsing broadcast TV materials with a view to developing a digital library of such materials where users will be able to access this stored archive, not just to select programs to stream and play, but to do other, content-based manipulations.

In coming years, people's home video watching habits and styles will become quite different from now, as new digital video content management systems and interactive TV are developed and start appearing on the marketplace. Because the material is in digital format, this will allow easy searching and browsing of video content information such as movies, TV broadcasts and personal archives. In the domain of non-temporal media, quite a few innovative and creative interface development efforts for browsing can be found including interfaces such as text

¹ The word Físchlár comes from fí (video) and chlár (programme) in the Irish language.

browsing/visualisation [2] [3] and the interfaces developed for browsing different types of data spaces [4] [5], but only very limited efforts and studies have been made on browsing interfaces to time-based media such as video, examples being [6], [7], [8] and [9].

Current efforts in digital video research are usually centred on automatic video indexing techniques while user-interface aspects are often demoted to the last stages of a project where some test users approve the acceptance of the system's interface, rather than the nature of video data and its possible end user interface and content visualisation being a separate subject of study on its own. It is indeed important to develop effective and efficient techniques to automatically index video materials, but unless we also consider how the output of these techniques can be properly used for providing good front end user interaction we may find that such technically-driven innovations will not be of any use in practice. What is needed is a more focused and systematic study that considers the characteristics of the video to be handled and to connect this to the development of a front-end information visualisation and user interface, with user-oriented, iterative evaluations and refinement.

In this paper we present our current work on keyframe-based video browsing user interfaces and user interaction by categorising and characterising different possible ways for users to browse video data. We use this idea to design innovative video browsing interfaces and conduct a focused study by adapting a parallel, iterative method with actively involving real users into everyday usage of a digital video system. In the following section we describe a digital video demonstration system called *Físchlár* which we use as our user interface testbed, then in section 3 we consider the nature of video data we are handling and identify three important dimensions in providing user interfaces using that data and some of the possible values for these dimensions. In section 4 we introduce *Físchlár* system's six distinctive video browsing interfaces and present a 3-dimensional space where we locate each of the six interfaces. Section 5 describes briefly each of the six interfaces with initial test users' comments and briefly what other possible interfaces could be implemented, and section 6 concludes with our plans for further interface development and evaluation.

2 The *Físchlár* System

The technologies needed to support manipulation of digital video information by computer have now developed enough to allow simple capture and playback of digital video on desktop machines. Beyond that, however, we are now at the point where video *content-based* browsing has become possible as new video indexing techniques have been studied and developed during the last few years. A fundamental primitive operation which is needed in order to allow any content-based operations on video is 'shot boundary detection' which automatically segments video material into individual camera shots by analysing differences of video content frame-by-frame, and then from each shot automatically select one or several frames ('keyframes') that represent the content of the shot. This generally results in a large set of pertinent keyframes as an output and this set of images is used by end users for browsing video content.

In the work of the Centre for Digital Video Processing at Dublin City University, we have been developing and evaluating effective shot boundary detection techniques [10] [11] [12] as a foundation for our fully working demonstration system, *Físchlár*.

Físchlár is a web-based system for capturing, storing, indexing and browsing broadcast TV material. It allows a user to set the recording of broadcast TV programmes by providing an easy-to-browse online TV schedule. Once the user sets the recording of a particular programme, Físchlár records and digitises the program when the TV broadcast time comes, using MPEG-1 format. The encoded programme is then fed into our shot boundary detection module and the resultant set of well-chosen keyframes are passed back to Físchlár to be used as part of a user's browsing interface. The user can then browse the keyframes with a web browser and can start playback at any point of the video by clicking on a keyframe.

The system allows users access to an archive or digital library of recorded TV programmes and has been running since July 1999. Through person-to-person introduction and active promotion within our department we now have our user base ranging from dedicated regular users to infrequent casual users, mostly postgraduate students and staff. The Físchlár system is described in more detail in [1].

While initially serving as a demonstrator for our shot boundary detection work, we are also using Físchlár as an underlying support system for further work in summarisation, personalisation and searching of video from an archive or library of stored programmes. For now, Físchlár has also become a testbed for various keyframe browsing user interfaces, although we plan to work on other video browsing methods such as skim [8] and others in near future. In keyframe-based browsing, once a basic set of keyframes are extracted from a video programme and stored in the system as metadata for that programme, it becomes simply a matter of how to present those keyframes onto the user's screen. Using the dimension framework that is described in following sections, we have been designing, implementing and testing several distinctive keyframe browsing interfaces all running on top of Físchlár. Although these distinctive interfaces are running independently, they use exactly the same set of keyframes from the video, which makes them more directly comparable.

3 Dealing with a Large Amount of Temporal Data

A set of keyframes extracted from video can be displayed in many different ways and styles. There have been numerous experimental and commercial digital video systems each with different visualisation and user interfaces, but a system having a single interface for video browsing would be too much of a simplification considering the many other possible ways that could have been useful, and would also lead to poor usage and frustrated users. In practice, a single all-encompassing interface would be impossible and several variations are needed, each of which has to tackle a well-defined specification which can be obtained from many sources such as potential exploitation of the video data itself and prospective future users' requirements. Acknowledging the various factors that could be considered, we focused on the exploitation of the video data itself, by analysing the characteristics of video data we have to handle. We identified different underlying *dimensions* the video data has and different *values* that each dimension can take. In other words, the video data we want to provide an interface to can be modelled in a multiple-dimensional space where each dimension contains a certain number of values. "Designing an interface" in this context becomes a matter of making a choice of different values for each dimension, different sets of choices resulting in different interfaces.

The data we manage is sets of large numbers of keyframes in temporal order. Dealing with a large number of keyframes, the first dimension we identify is what we call "layeredness" - a matter on whether to display all available keyframes directly to users or to allow a way for more selectively presenting them. The temporal nature of the data leads us to think of two more dimensions - "Provision of temporal orientation," or how to make time information of video visible to the user, and visualisation style we call "Spatial vs. Temporal visualisation."

These three dimensions and their possible values will be considered now.

3.1 Layeredness

A set of keyframes selected from shot boundary detection more often than not means a very large number of keyframes. For example, in our experience a 30 minute soap opera might generate more than 400 keyframes, and a large number of keyframes can be impractical for browsing. The method of reducing the number of keyframes ("keyframe filtering") may vary from simply selecting an arbitrary keyframe by grouping all keyframes in a regular interval, or more complex ways of content-based and semantic-level keyframe grouping.

Whatever the method of reducing the number of selected keyframes may be, this implies a layered presentation where several different granularity levels can be chosen as an option.

At one end of the layeredness of keyframe-based presentation there will be a single keyframe that represents the whole video clip, as is frequently done for displaying search results for a video clip in experimental systems such as [13] [14] [8]. One cannot expect this single keyframe to be able to convey the content of the whole video, but still it works as a kind of icon for the video and locates clips within the whole video. At the other end of the layeredness scale there will be the most detailed, full listing of keyframes with each representing a camera shot in the video, as most of current semi-automatic video cataloguing systems do [15] [16] [17]. Whereas this full keyframes listing shows the most detailed view of the video content, it loses some of its purpose for browsing because it takes so long to look at all the keyframes.

Between these two extreme layers, we can think of intermediate layers with different levels of detailed. It will not be possible to say which level of granularity is best for every situation as one user in a situation will have different needs from another user in the same or a different situation - a busy user might want to look at what the video content is on one screenful of small keyframes, while a user looking for a particular scene might want to browse in as detailed a way as possible and may be willing to spend more time, and so on. It might also be the case that different types of user groups prefer browsing at different levels of granularity, leading us to consider a browsing interface with the level of granularity customisable for users. In general a user might want to have an overview look of the content and if s/he finds it interesting then may want to see a more detailed view. This idea of "overview first, zoom and filter, then details on demand" [18] has been amply emphasised in the field of information visualisation, and should be desirable to apply in keyframe-based interfaces as well.

The important values we identified in providing layered keyframe presentations are:

- The number of different layers available
- The navigational link between different layers (and visualisation to show which layer the user is currently in)

The above ideas can be summarised as the following figure:

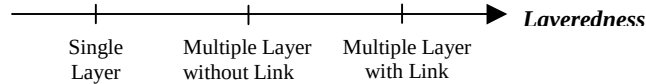


Fig. 1. Layeredness dimension and some values

3.2 The Provision of Temporal Orientation

With the video medium having a temporal nature as one of its fundamental characteristics, it is important to provide a cue for the sense of time in user browsing. A set of keyframes usually provides no proper time sense or temporal orientation, other than a rough progression from one keyframe to the next. Looking at the keyframes selected based on camera shots can even distort and mislead the time orientation, as the number of shots and thus keyframes from one time segment can be very different from those in another segment of the video but of the same length, depending upon the frequency of camera shot cuts in the video content. A familiar time orientation used in many interfaces is known as the 'timeline bar' (either as an indicator or controller) and is usually found in most video player software such as Microsoft's Media Player or RealNetworks RealPlayer. In keyframe browsing, it is easy for the user to lose the sense of time and at what point s/he is currently looking at in the video, and in our work we considered some way of indicating the time sense as an important dimension in browsing. Several existing examples of this implementation are found such as simply time-stamping the numeric time in each keyframe as in [19] [16] [15], displaying quantifiable objects to visualise shot length [20] or a timeline bar indicating current position of user's keyframe browsing such as in [17]. We can identify some of the possible values for this dimension, as:

- The indication of absolute time in the currently viewed keyframe or set of keyframes (exactly how much into the video the current keyframe is)
- An indication of the current browsing point in relation to the whole video

The above ideas can be summarised by the following figure:

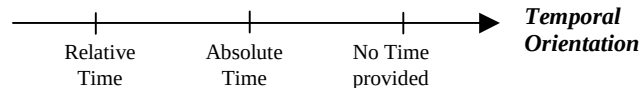


Fig. 2. Temporal Orientation dimension and some values

3.3 Spatial vs. Temporal Visualisation

The conventional keyframe browsing idea would be to miniaturise each keyframe into a small thumbnail size image and present many of them on one screen - much the same as the digital image retrieval user interfaces used in art gallery booths or CD-ROM picture collections where the user's query result is displayed as a set of

thumbnail images so that s/he can quickly browse and identify the images s/he wants. But it is also possible to present keyframes on a user's screen one by one, the system automatically flipping through the keyframes. Considering the fact that the original video is time-based and sequential, it seems to make sense to present keyframes in an identical time-based and sequential manner, and this might even provide the feeling of watching the video. Current digital video systems' user interfaces based on keyframes mostly feature showing thumbnail size keyframes spatially such as MediaArchive [19], FRANK [21], MediaSite Publisher [17], Excalibur Screening Room [16], VICAR [22], Virage VideoLogger [15] AT&TV [23], and infrequently found are the systems presenting keyframes temporally such as DVB-VCR [24], WebSEEk [25] and movie browser tool [26]. Such an unusual video browsing study as comparing spatial presentation with temporal presentation in terms of different user tasks is also found [27] and our study in a sense extends this work by including more variety of visualisation styles. Then again, this can be summarised as the following figure:

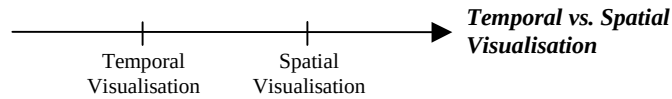


Fig. 3. Temporal vs. Spatial Visualisation dimension and its two values

3.4 Using the 3-D Representation of Video Data to Specify Browsing Interfaces

Assembling these analyses of video data into one leads to a 3-dimensional space where each axis represents each dimension drawn above. In this space we can locate positions of several existing video browsing interfaces, for example SWIM hierarchical browser [28], DVB-VCR [24] and AT&TV [23]. (see Figure 4 below).

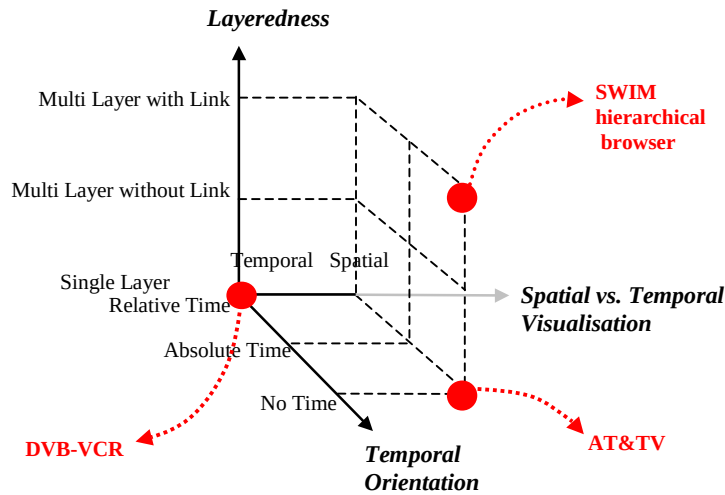


Fig. 4. Dimension Space and three example interfaces' positions

This way we can see different browsing interfaces more globally in their relative positions in terms of three dimensions. All three example interfaces located in this space are in fact choices among alternative ways of designing interfaces. Our approach is to propose several further well-specified browsing interfaces in this space, by locating a particular point in the space. Each of our proposed interfaces aims to fulfill the identified values from each dimension. As the next section will illustrate, we designed and implemented six distinct browsing interfaces for the Físchlár system, which provide good coverage of the dimension space.

4 Dimension Space and Six Físchlár Video Browsers

The Físchlár system is currently running with six distinctive keyframe browsing interfaces. Our users are free to use any of these interfaces to browse and play TV programmes they, or others, have recorded. This section will address the six interfaces' individual specifications in the dimension space, and the next section will describe each of the six interfaces. As the following figures will show, our objective is to cover as well as possible the previously defined dimension space in order to propose alternative browsing possibilities for the Físchlár system. The six Físchlár video browsing interfaces are listed below and are attached to a specific colour which will be used consistently in the following figures throughout this paper.

Table 1. Físchlár Video Browsers and assigned colours

	Scroll Bar Browser (SB)		Overview/Detail Browser (OD)
	Slide Show Browser (SS)		Dynamic Overview Browser (DO)
	Timeline Bar Browser (TB)		Hierarchical Browser (H)

4.1 Layeredness and Físchlár Interfaces

The idea of layeredness can also be found in text indexing in information retrieval systems. In those systems, there exist alternatives from the classical list of keywords to complex indexing languages. Choosing an alternative is mostly a problem of properly identifying the users of the information retrieval system: the more knowledgeable the user is about the domain, the more complex the indexing language has to be to satisfy him/her. The layeredness in our dimension space can be viewed in the same way: from a simple, single layer keyframe list to a more complex, multiple layer list. The more knowledgeable the user is about the video content, the more layers and complex arrangements of data s/he will want to use.

Our choice here is to design three single layer interfaces and another three multiple layer ones (for the latter, two that provide point-to-point navigational link among layers, and one without any navigational link).

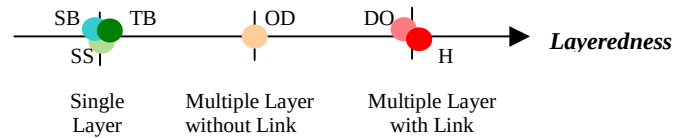


Fig. 5. Six Físchlár interfaces on Layeredness dimension

4.2 The Provision of Temporal Orientation and Físchlár Interfaces

Providing time information (either absolute time into the video or displaying the current point of video in relation to the whole length) or not are possibilities that must exist as alternatives to each interface. Therefore we decided to widespread our six interfaces in all the values. Of the six Físchlár interfaces,

- two interfaces do not provide time at all;
- two interfaces show the current point relative to the whole video;
- two interfaces show both the current point relatively as well as absolute time

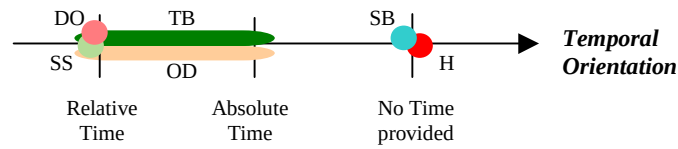


Fig. 6. Six Físchlár interfaces on Temporal Orientation dimension

As can be seen, the distribution of the six interfaces in terms of this dimension has been made so that different interfaces provide properly covered combinations with the layeredness dimension - single layer with time, multiple layer with time, and so on.

4.3 Spatial vs. Temporal Visualisation and Físchlár Interfaces

The choice between spatial and temporal visualisation of keyframes might be largely a matter of the available time of the user. We observed that often the initial test users preferred spatial visualisation unless the video is quite short. This may explain why we have mostly implemented spatially displayed interfaces instead of temporal ones. From this idea we introduced another interface that mixes both spatial and temporal visualisations, which covers both values in the following figure (Dynamic Overview Browser).

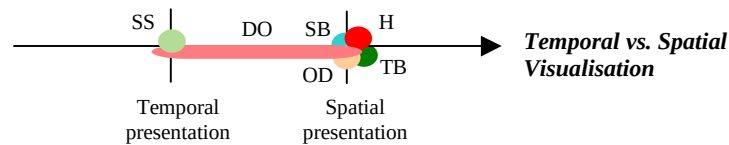


Fig. 7. Six Físchlár interfaces on Temporal vs. Spatial Visualisation dimension

4.4 The Six Físchlár Interfaces in Dimension Space

The above three figures can be combined into one 3 dimensional space with one axis representing each dimension. This combined figure shows all essential dimension value choices of the six Físchlár browsing interfaces at one glance.

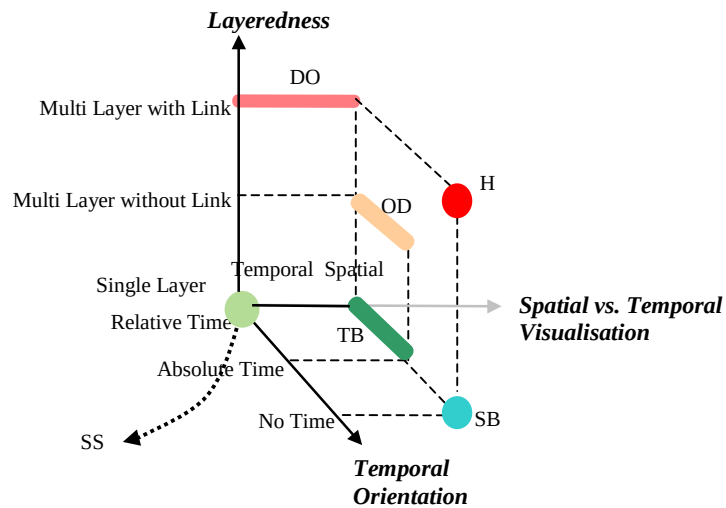


Fig. 8. Dimension space and current six Físchlár interfaces

Following this positioning of our six browsing interfaces in the dimension space presented, we now go into details of each interface.

5 Físchlár Video Browsing Interfaces and Characteristics

This section describes each of the six Físchlár interfaces which were introduced and viewed in the dimension space above. Presented below are screen shots, short descriptions and some selective and indicative comments from our users. Beginning with the Timeline Bar front end, interaction and dialogue with our users brought up new ideas and critiques, which then led to further ideas and then into additional front end interfaces. Added interfaces were again tested with users and, over time, some obvious usability problems were identified and removed as we then started to stabilise the features of each user interface.

The user interfaces presented below are a snapshot of the on-going refinement and discussion with our users within the Lab so far and will be further refined, although we note that these six user interfaces are fairly stabilised and fixed at this point.

5.1 Six Físchlár Browsing Interfaces

Single Layer	Without Time		Scroll Bar Browser: The user scrolls up and down using the scroll bar to the right to browse all the keyframes for the video arranged next to each other. Familiarity with normal scroll bar, easy control and immediate response are strong points. Some user comments: Simple to use; I feel pictures are more accessible; too much information; doesn't look organised.
	With Temporal Orientation	 	Slide Show Browser: Keyframes are automatically flipped through one by one, and at the bottom scroll bar the timeline indicator shows the current point in the video. Easy to lose the context as the user can only look at one keyframe at a time. Some user comments: It takes too long to see all the keyframes; it would be better to be able to control the speed of flipping. Timeline Bar Browser: Presents fixed number of keyframes (24) on one screen, and as the user moves the mouse cursor over the timeline bar at the top, the screen of keyframes flips through to the next set of keyframes. The timeline bar provides temporal orientation, windows are proportional to the time covered by the set of 24 keyframes, and the ToolTip shows the exact time of the current screen. Some user comments: Browsing screen is very inviting; easy and simple.
Multiple Layer	Without Navigational Link		Overview/Detail Browser: Initial screen displays a small number of "significant" keyframes selected from throughout the video, thus providing an overview. When the user wants more detailed keyframes, s/he clicks a button at the top and then the browser changes to the Timeline Bar Browser. Some user comments: Not sure if overview screen is necessary at all.
	With Navigational Link	 	Dynamic Overview Browser: Initial screen displays one small set of keyframes (as an overview), then when the user brings the mouse to any of the keyframes, the pointed keyframe slot starts flipping through the detailed keyframes within that slot. Small timeline bar appears under each keyframe, indicating current point. There is no sudden major screen change that can disorient a user, and overview and detail are linked to support navigation. Combined spatial and temporal browsing. Some user comments: Very novel; It takes too long to see each slot's keyframes. Hierarchical Browser: Initial screen displays 6 keyframes at the top level, then as the user brings mouse cursor over one of the keyframes, another 6 keyframes(lower level) within selected keyframe are immediately displayed. Thus by using the mouse the user can browse the hierarchically broken down keyframe structure in a highly interactive way. This idea has been developed and implemented earlier [6] [28]. Some user comments: I'd expect the structure precisely broken down by topic.

5.2 Further Improvements and New UI Additions

The dimension space is useful for us to come up with new browsing interfaces. For example, in the space we can think of a new interface by dragging the “Scroll Bar Browser” inward along the *Temporal Orientation* axis to provide time information - this will be a Scroll Bar Browser variation with stronger time orientation for the user.

Also, we want to drag all the bottom user interfaces up to the top on the *Layeredness* axis by implementing additional layers to Scroll Bar Browser, Slide Show Browser, Timeline Bar Browser and Overview/Detail Browser.

Figure 9 below shows some possible interesting browsers in the dimension space.

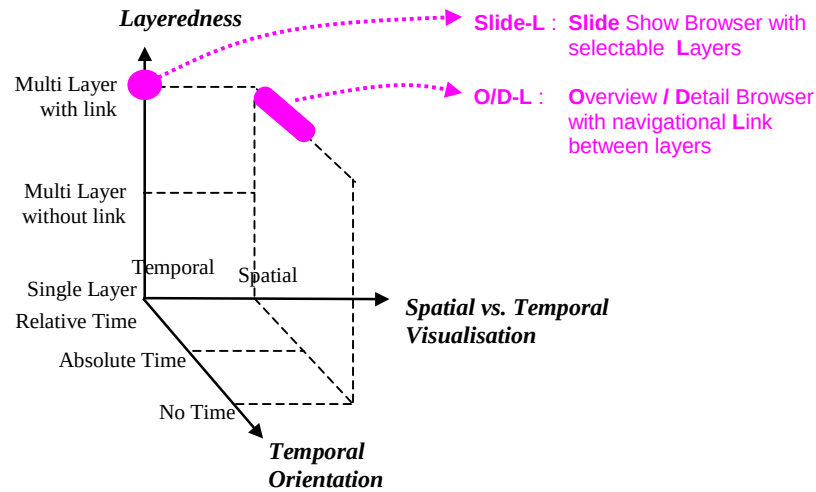


Fig. 9. Dimension space and some possible interfaces

A final note is that care has to be taken when shaping a new browser idea and implementing it. When a new browser is designed and implemented, it might be the case that the general usability problems hinder the browser and make it a bad browser when the novel idea behind it might be a very good one and worth consideration. One should not hurriedly abandon the new implementation and bury a novel browsing idea just because some artefacts make it a poor browser overall. The designer has to carefully and creatively rearrange the interface in such a way that the usability problems can be avoided or minimised while still keeping the original browsing idea.

6 Conclusion

In this paper we described our work in developing video content browsing interfaces based on keyframes. We identified three important dimensions in video data browsing, then used them to implement some of the representative interfaces. While we continue to test these interfaces with users to improve them, we also add completely new interfaces by identifying what the current interfaces are lacking in the

dimension space and again getting users to try them. This user-oriented, parallel and iterative design method we find very suitable for developing novel interfaces such as video content browsing, and Físchlár system has been serving this purpose very well.

We also find a need to more specifically categorise users' browsing tasks depending on a person's cognitive direction in browsing, rather than covering it all as simply 'browsing'. A user might be browsing to get the overall gist of the whole programme as quickly as possible, or to find a certain scene s/he had already seen in order to watch that particular moment again, or even to find out why's or who's by browsing here and there to gather related information in the video. For certain kinds of tasks we plan to record all episodes of a TV programme (for example a popular soap) and use this as a testbed for the users who have already seen the series before and want to find out particular scene or scenes in the recording, so that we can see a particular browser being better utilised for content-knowledgeable users, and so on.

Acknowledgements: The authors acknowledge the help given by Kieran Mc Donald in implementing some of these interfaces.

References

1. Lee, H., Smeaton, A., O'Toole, C., Murphy, N., Marlow, S., O'Connor, N.: The Físchlár Digital Video Recording, Analysis, and Browsing System. *RIAO '2000: Content-Based Multimedia Information Access* (2000) 1390-1399.
2. Hearst, M.: TileBars: Visualization of Term Distribution Information in Full Text Information Access. *Proceedings of ACM SIGCHI '95* (1995) 59-66.
3. Veerasamy, A., Belkin, N.: Evaluation of a tool for visualization of information retrieval results. *Proceedings of ACM SIGIR '96* (1996) 85-92.
4. Shneiderman, B.: The eyes have it: A task by data type taxonomy for information visualizations. *IEEE Conference on Visual Languages* (1996) 336-343.
5. Chi, E., Riedl, J.: An Operator Interaction Framework for Visualization Systems. *Symposium on Information Visualization (InfoVis '98)* (1998) 63-70.
6. Mills, M., Cohen, J., Wong, Y-Y.: A Magnifier tool for video data. *Proceedings of ACM CHI '92* (1992) 93-98.
7. Aigrain, P., Joly, P., Lepain, P., Longueville, V.: Representation-based user interfaces for the audiovisual library of year 2000. *Proceedings of SPIE, 2417, Multimedia Computing and Networking* (1995) 35-45.
8. Christel, M., Winkler, D., Taylor, C.: Multimedia abstractions for a digital video library. *Proceedings of ACM DL '97* (1997) 21-29.
9. Boreczky, J., Girgensohn, A., Golovchinsky, G., Uchihashi, S.: An Interactive Comic Book Presentation for Exploring Video. *Proceedings of ACM CHI '2000* (2000) 185-192.
10. O'Toole, C., Smeaton, A., Murphy, N., Marlow, S.: Evaluation of Automatic Shot Boundary Detection on a Large Video Test Suite. *The Challenge of Image Retrieval (CIR 99) - 2nd UK Conference on Image Retrieval* (1999).
11. Smeaton, A., Gilvarry, J., Gormley, G., Tobin, B., Marlow, S., Murphy, M.: An Evaluation of Alternative Techniques for Automatic Detection of Shot Boundaries in Digital Video. *Irish Machine Vision and Image Processing Conference (IMVIP 99)*. Dublin, Ireland, 8-9 September 1999.

12. Browne, P., Smeaton, A., Berrut, C., Murphy, N., Marlow, S., O'Connor, N.: Evaluating and Combining Digital Video Shot boundary Detection Algorithms. *Irish Machine Vision and Image Processing Conference (IMVIP 2000)*, Queens University, Belfast, Northern Ireland, 2000.
13. Compton, C., Bosco, P.: Internet CNN NEWSROOM: A digital video news magazine and library. *IEEE International Conference on Multimedia Computing and Systems (ICMCS 95)* (1995) 296-301.
14. Li, W., Gauch, S., Gauch, J., Pua, K.: VISION: a digital video library. *Proceedings of ACM DL '96* (1996) 19-27.
15. Virage Inc. *VideoLogger 4.0*. <http://www.virage.com/products/videologger.html> (last visited 26 April 2000)
16. Excalibur Technologies Corporation. *Excalibur Screening Room*. <http://www.excalib.com/products/sr/index.shtml> (last visited 26 April 2000)
17. MediaSite Inc. *MediaSite Publisher 4.0*. <http://www.mediasite.net> (last visited 26 April 2000)
18. Shneiderman, B.: *Designing the user interface: strategies for effective human-computer interaction*. 3rd Ed. Reading, MA: Addison Wesley Longman (1998).
19. Euromedia Project - Distributed Multimedia Archives for Cooperative Production. <http://www.foyer.de/euromedia> (last visited 26 April 2000)
20. Arman, F., Depommier, R., Hsu, A., Chiu, M-Y.: Content-based browsing of video sequences. *Proceedings of ACM Multimedia '94* (1994) 97-103.
21. Simpson-Young, B., Yap, K.: FRANK: trialing a system for remote navigation of film archives. *SPIE International Symposium on Voice, Video and Data Communications* (1996).
22. VICAR Project - Video Indexing, Classification, Annotation and Retrieval. <http://iis.joanneum.ac.at/vicar/> (last visited 26 April 2000)
23. Mills, T., Pye, D., Hollinghurst, N., Wood, K.: AT&TV: Broadcast Television and Radio Retrieval. *RIAO '2000 - Content-based Multimedia Information Access* (2000) 1135-1144.
24. DVB-VCR. In: *SMASH (Storage for Multimedia Applications Systems in the Home) Project final report* (1998).
25. Smith, J.: Searching for images and videos on the World-Wide Web. *Technical report #459-96-25*, Center for Telecommunications Research, Columbia University (1996).
26. Marchionini, G., Plaisant, C., Komlodi, A. Interfaces and Tools for the Library of Congress National Digital Library Program. *Information Processing & Management* 34(5) (1998) 535-555.
27. Tse, T., Vegh, S., Marchionini, G., Shneiderman, B.: An Exploratory Study of Video Browsing User Interface Designs and Research Methodologies: Effectiveness in Information Seeking Tasks . *Proceedings of Annual Meeting of the American Society for Information Science (ASIS'99)* (1999) 681-692.
28. Zhang, H., Low, C., Smoliar, S., Wu, J.: Video parsing, retrieval and browsing: an integrated and content-based solution. *Proceedings of ACM Multimedia '95* (1995) 503-512.

Functional and Intentional Limitations of Interactivity on Content Indexing Topics: Possible Uses of Automatic Classification and Contents Extraction Systems, in Order to Create Digital Libraries Databases

Dr. Florent Pasquier

Associate Professor, IUFM of Paris
21 rue du Pere Guerin
75013 Paris FRANCE
pasquierf@paris.iufm.fr

Abstract. The context of the creation of self-learning contents knows a revolution by the use of digital tools. The process to realize video was based previously on the use of analog tools. It moves nowadays towards digital ones: studios are equipped to work with digital data and digital materials under the pressure of international standards and components. On the same time, capture and storage pictures by videotapes evolve towards hybrid supports (Digital/Video), which will become entirely digital soon (computers hard-discs and networks). The current creation process which uses a kind of simulated interactivity, waits new materials and synopsis that will allow a true and complete interactivity. This interactivity needs to be used through a full digital multimedia way. The current evolution of the market digital systems seems to empower the role of the creator and the designer. They might start their creation with previous and genuine videos, by exploiting the ones which suit the best what they need, picking them up on video databases. But the problem is to be able to locate the video wanted. Many systems of automatic recognition (voice, pictures, videos...) are already functional to help them in this heavy task. But the question of using them in a relevant context, as the educational one per example, is still not resolved.

Introduction

In the future, André VITALIS (1994, p.42) says that "the information's digitalization is able to make easier and greater the TV viewer interventions into the choice and development of the TV programs". But this needs an increase of the stored and broadcasted videos, which require especially evolutions regarding the automatization of contents analysis for the creation of catalogues, and also the creation of special browsers to play these documents. These operations are essential as data's treatment should allow to extract clearly from databases the specific topics to be proposed by providers and broadcasters to end users (Resources Centers, digital libraries...). But such project has to become fully automatic, because manual treatment is not efficient to cover the whole audio-visual production. So, a state of the art of automatic recognition techniques is obviously necessary.

The focus of our work¹ is to analyze techniques for digital video systems and particularly their applications in the field of education.

About technical aspects to evaluate possible resource center uses, we digitized two video products made for language learning (News Busters and Video Presentation), in MPEG1 norm. Each method proposes a genuine document of TV-news (from 3 to 5 minutes) and then completes it with didactic sequences (30 to 45 minutes). We indexed these linear movies in respect to the sequences, which compose them, in order to allow for the users several entrances into the explanations. Such concept creation was possible because of the independence of the sequences composing these movies. Authors had placed sequences in a determined order, but they were easily removable. The interest of those video was to simulate a kind of interactivity: an explanation grid based on the answers to the questions WHO, WHAT, WHERE, HOW, for the first method, and a specific order of the sequences corresponding to the student's foreign language level, in the second one.

This indexation was "hand made", and so it had been tedious and long. It has rapidly shown its limits in fineness as well, as the more you are looking for a detailed level of the analyze, the more it becomes hard and long to catalogue and to index.

Worst, navigation screens and three walk are moreover complex.

What Kind of Articulation for Video Sequences?

In this way we could propose to the viewer to sort out automatically the videos according to their themes (documentary, fiction, animation...), or their format (2 minutes, 15 minutes...). The point would become to associate different types and circulation levels into the same video database (several navigational strategies), according to the user targets : limited consulting time, search for special topics... The viewing wouldn't be fixed once but, in the contrary, should change and become paramettable. Then, the result will be enhanced video product that should allow the viewer to select specific movies among a digital video library in creating automatically a personalized viewing.

Already, some products propose original consultation modes :

Jean-Paul MAZEAU (1992) designed a system allowing, after carrying up virtually recorded pictures on varied supports, to exploit them sequentially or in parallel, by introducing iconic data concepts and by using rules managing every relation between the contains to learn and the pictures that help the comprehension, without any limit of how many video plans neither of the complexity of the relations between those plans.

Another prototype have been designed in the early 1994 by Gilberte HOUBART at the MIT Media Lab (Cambridge, USA), " Viewpoints on Demand ", included in a research program about the "Daily news of the future". The system proposes to the

¹ "Interactive video to enhance oral English comprehension. Designing and use of a digital video server", December 1997, Paris 7 Denis Diderot University.

user to select one or several journalist-reporters, theme titles or arguments, and a watching length (5 to 20 minutes). It uses as starting point a video report base about the Gulf War : four video interviews during two or three hours. The result deals between author and spectator leads to a video sequence generated by the system, in respect of the choices and creating special links. Reports have been previously reassembled into chapters by the author, allowing linking them automatically afterwards. The target is to be able to make personal video editing on demand. Among the critiques heard after a presentation², the false contains' objectivity and the lack of fine analyze were reproached. By another way, the rebuilt of a linear program from a multidimensional database seems to be a regression to traditional linear watching mode, when it seemed to be possible to obtain something new.

By the way, Pierre MOEGLIN (1996, p.36) observes that "sequences of which the unfolding can't be interrupted before the end (...) impose a specific consultation order, while some others fine informations can be understood only by transversal reading and by relation (more in a telecommunications logical way). Too many continuation solutions handicap the multimedia acquisition ". Switching from a linear to a fluctuating consultation is not sufficient enough.

According Daniel COSTE (1996, p.45), " there is a point to wonder whether the conditions to fulfill those navigations are not to be only for entertainment or random navigation, but answer truly apprenticeship trajectories determined in purpose.

In order to know better the documents, and then use them better, researchers often chose a descriptor analyze type methodology. Carmen COMPTE (1993a, p.26, 29,62) presents for example a non exhaustive list of the elements that compose a film : principal theme, sub-theme, verbal kind, language and accents, linguistic lacks and illustrated speech act, background elements, stereotypes, format and kind of the emission, visual indications for comprehension (titles, photographs, drawings, logos...), narrative type (figurative, verbal, musical)...

As an example she propose (1993a, p.50-51) a detailed analyze of a TV-news program in France (Antenne 2, 8 PM, 1982). "Close ups aim the essential elements for the understanding of the subject (...). Large plans allow the localization (...). In a social information report, they are more numerous than the close ups which are more appropriated for personal subjects. The medium plans and close ups, are more likely used for causing an impact on the spectator, the nearing of the subject implicate him much more...(..).

The technical analyze acknowledges the understanding of visual writing conventional rules. Even limited, it assures a global minimal understanding and allows, in the same time, the possibility for the learner to focus on the critical details that the teacher can then explain³, in showing how " to unbuild " the picture (p.61).

A typology of TV reports is proposed (p.59), from those specially realized for a special event, passing by those bought to foreign televisions, and those using archives

² NOV'ART Special magazine, February 1996, Paris, France.

³ For a documentary's analysis (p.51), the interpretation will be different, every format using its own specificities, serving special objects : classically or ideally , the report is subjective, argumentative, actual and original while the documentary is objective, descriptive, global and exhaustive.

pictures reused for new reports. At one higher analyze level, some emissions' type structure are enlightened: in star type (usual TV news), linear sequence (with or without transitions : film, sitcom, series...), interlaced (soap opera), in spiral (ground magazine) or in parallel (advertising).

Pierre MASSELOT (1988) also proposes an analyze grid specifically adapted to the 8 PM TV-news report on Antenne 2 during the year 1985. The target is to isolate constitutive elements and functions in order to "constitute a practical trace of the different parameters operating into the bulletin: temporal data (succession of the subjects, length, spot of the subjects on the videotape meter...), visual data (pictures at the screen), scriptural data and their place (typography...), sound data (locutor identified in off-voice or not, linguistics - the verbal-, non linguistics - noises and music-).

We obtain a multicode tubular document which is supposed to design schematizing, decoding/transcoding, and temporal structure operations". A fine "macrostructure" is then transcribed into a general grid including the following elements: length, theme, visual (screen, scriptural), audio. The author starts with an analogy between the written newspapers and the TV bulletin successive screens, leading to precise that the objects will be analyzed on a base of screens. Those TV screens are of two sorts : full screen or exploded screen, the last one decomposed into "Presenter plus mention" or "Presenter plus figure(s)". The succession between screens can be then analyzed, in order to create "visual trajectories".

He conclude by saying that "like every multicode message, TV bulletin's decipher suppose a lots of operations converging to a better understanding of the TV informative product (...)"

If this "macrostructure" seems interesting to us and fully exploits the techniques of this period, we now have the possibility, thanks to digital technologies, to go further in the analysis to spot and isolate informations until the word itself, as the more tiny sense unity.

Those analysis applied to audiovisual documents proposed in a database, would allow a multi-criterion quickness research and access to numerous examples in order to adapt to multiple possibilities (use in course or in auto-formation), as well in a language context (we propose a linguistic multi-indexation: forms -length, future...-, didactic -what usually meets problem in the study: accents, idiom words, frequent natives' mistakes...-), as well for didactic or picture analysis.

But the main problem of the previous propositions dwells into this capture case by case of the multiple sound and picture contained on the movies, followed by their ranging and indexing into the grids. It's expressed Véronique PUGIBET (1996, p.55) : "the database and multimedia resources allows the learner to "dive" directly into those kind of sign mosaic of the pictures, sounds and texts, but previously "organized", "didactized" for him to suit it to the maximum. It is important to underline that, through all the experiments presented, the amount of work required for multimedia sequences' development, seems colossal". Some experiences show that those automatic procedure tasks could be possible to carry out this work faster, and to reach a sufficient information volume to answer efficiently to any content request, and justify its employment several contexts.

The Automatic Picture Analysis

Examples do not lack in this field: recognition of the form defaults on manufactured products (Seddix HADFI, 1991) by exploiting binary pictures or by grey level, with the objects outlines put in correlation with form detection algorithms (PREWITT, SOBEL, CANNY, KIRSCH, LAPLACE's algorithms..) ; animals (or other objects) recognition methods (Pierre CLEMENT and Valdiodio NDIAYE, 1995) ; automatic detection and localization of the front or profile faces (Renaud SEGUIER, 1995) by modelization of the elliptic curved forms, in real time and whatever the characteristics of background or light are...

But we notice those researches are all very specialized: medicine, biology, kinetic, meteorology...

Should a global analysis be possible? Actually it is not, but we could distinguish successive treatment phases of such work, and then assemble the results. If, as wrote so Pierre BOURDIEU (1996), "in fact, paradoxically, the picture's world is dominated by words", then would it be possible to automatize, like for the writing, and organize a reading that would follow this structure ? But, before coming on this question again, let's observe how a fixed pictures string behave, giving an illusion of motion.

Automatic Pictures Sequences Recognition

Julia KRISTEVA notice (1981, p.312-314) that at its beginning, cinema sees itself as a new language that seeks its syntax and its movie assertion even more willingly than, mute, it was looking for a different structure from the one of the words. The main differences are operating, according to Christian METZ (1968) in the way that " the picture is always the word, but never reach a language unity" . There is so a language founded on a syntactic base. The most efficient way (I mean the most realistic) to systematize an automatic analysis of the audiovisual creation is to keep a single approach (semiotic and denotative) based on the technical filming possibilities (camera travelling...) and editing choices (linking type of the sequences).

Mourad CHERFAOUI (1995) presents some video indexing possibilities similar to those of structured texts, from an interpretation and a structure of the original documents, according to the visual and sound contents, which aims to allow multiple travels. Special marks must allow an easy and personalized reading of the video writing. The visual extraction of pictures is made after a sequence of picture analysis. This process is automatic (which avoid a tedious manual treatment) by plans ruptures research and camera traveling recognition, and their accords with predefined models, which allows multiple interactions possibilities.

A practical application of these principles exists with Videoscribe product, used at the French National Institute of Audiovisual (INA). The paper specifies that it is able to describe audiovisual documents at several levels of the structure, from a macroscopic documentary management to a detailed analyze at the sequence level, by the automatic detection of plans ruptures and dynamic audio signal analyze. These

informations allow navigating on a database represented by pictures board corpus describing the recognized movies segments, and to access easily to the chosen plans sequenced on videotapes.

In 1996, Nabil MADRANE, from those analysis problems and the difficulty of representing possible navigations into video documents, develops a scene analyze tool on a model which, from pictures analysis, detect sequences and dialogues, index contains in a macro-segmented graph, leading to a visual compact description, with actors identification, and go further a static representation as it also includes camera travelling, plans occurrences and motion within the frame.

The same year, Brigitte SIMONNOT, resume this representation in a conceptual graph which represent complex contains and nuances connected with documentation and video media fields. Those graphs support queries by key words, interpreted by intelligent agents. They lean on the user's judgments to improve the research and likely to evaluate their performances, in order to evolve in a dynamic and efficient mode in cooperation between them.

Automatic Sound Recognition

Software as the one used by LAVAC system (Anthony STENTON, 1992) is able to find automatically the separations between words in a phrase, and distinguish into oral production as much units as wanted, in settling the tolerance by the sound lack length⁴. The difficulty would be then to range them automatically according to their characteristics into a grid.

We have not found another automatic system allowing such kind of function. Anyway, the main difficulty is still, according to Maryse QUERE (1991, p.40), that automatic treatment of natural languages is willing to be very complex, as one can judge in verifying the lacks (...) in continuous word comprehension, independently of the locuter or in automatic process translation. Such activity requires effectively the collaboration of very different disciplines such as calculator linguistic, general linguistic, contrastive linguistic, computing, formal logic, cognitive psychology ... ", and also (Pierre LOARER and Etienne NORMIER, 1996, p.115) " automatic linguistics and statistics ".

Some researches had paid attention to the lips shape into the sound production. A systematization aiming an automatic recognition was seeked by Mohamed-Tahar LALLOUACHE (1991), who studied the vocal conduit movements, and especially the lips shape (labiometry), by automatic analyze of video sequences in color.

Another solution could be to have an intermediate step by the writing of the oral production previously recognized.

⁴ Sound needs to be of .wav type file, which exclude an automatic treatment of MPEG type audiovisual files, unless previously dissociating the video and audio sub-files composing the MPEG file.

Does not Bernard STIEGLER (1985, p.78) justify a successive approach of the interpretations when he recalls that "Jacques DERRIDA wrote that the word is always already written "? This writing being interpretable by linguistic analysis, which "resolves the oral speech in an ended series of basic information units " (JAKOBSON, 1965, p.87).

Automatic Texts Recognition

Artificial intelligence process is the field where researchers' works have particularly been improved during the 90's. As the vocal dictations in software are progressing steadily in becoming soon able to write a normal speed pronounced sentence⁵, we can imagine the next process. In the first step, a process would transcribe the audio content of the video (by automatic recognition and transcribing). In the next step, this production would be analyzed by specific softwares in order to find out the meaning and to complete analysis grids.

The works based on the text analysis operate in different ways. According to Maryse QUERE (1991, p.40), "artificial intelligence inspired techniques (...) are solely able to take on account the field in the whole diversity and in the extremely delicate hierarchy of the different items evolving in the texts determination".

An example is given by François CHARPIN (1995, p.111-124), with the word morphologic analysis, "considered character strings delimited by blanks". It is decomposed into "two main operations : in the first step, with aid of grammatical rules, to set the inventory of all the vocabulary forms (lemmas) able to provide this word ; in the second step, to identify, within those lemmas , the one(s) recognized by the dictionary. Whether it is a processing device or a human learner, the process remain the same. It lays on the precise grammatical and lexis knowledge suitable to a given idiom (...). The pedagogic expertise lays on the distinction between the rules knowledge and the lexical knowledge".

According to Maryse QUERE (Ibid.), " What is important is to apply the rules (and not to be linked to a random shape system). Expert systems, in this way, improve significance by separating carefully the data and the treatments which are made on them, they allow to design and interpret linguistics objects, in respect of language structure".

Conclusion

There is most than a single step left to switch from analog to digital interactive and interactive video. The trouble dwells into the speech logical and the mainly linear statement that we keep used until now with video analog device. This difficulty as

⁵ At the beginning of 1996, 01 informatique n° 1392 p.36 presented the IBM's Personal Dictation supposed to have a 97% voice recognition rate.

been surmounted by only few designers or artists⁶. In an educational context, what are the possible approaches and how to give the control to the end user? The National Council For Educational Technology (UK) proposes the following approach⁷ and provide us some conclusions :

"The non-linear media of IV, CD-I, CDTV and CD-ROM offer a fundamentally different way of engaging with information, from the linear media of books and television. User control means that there is no single 'author' who governs the sequence and pacing of the material - the information experienced is a genuine interaction between user and producer. Control is important because it affects both pace and sequence, and both are important for learning to be effective. We saw many instances of pupils taking advantage of these features in a way that was likely to benefit their understanding (...). CD-I can both support close analysis of visual data, and can offer a form of illustrative feedback (...). Teachers expected the IV to be more interactive than it was, providing more feedback and guidance. This was the feature in which the discs fared worst, in their judgment".

Since, researches tend to turn towards the increase of technical performances (Christophe DUTEIN, 1994), and to the documentary needs analysis, such as automatic indexation of video sequences, more likely than conceptual indexation or needs of end users.

In order to set the video picture interactive itself, a congress at BRISTOL (UK) displayed tools defining sensitive zones on video objects in motion (by themselves in fixed plan or by the camera travelling). In that way, the MOVVIDEO product advertisement's tablet indicates that its goal is to reproduce the hypertext's philosophy on a motion video. Wherever on the video you are, click on the object or actor for whom you wish extended informations. Then you will get informations about him on text, photograph, sound, video and even hypervideo forms (...). Sail into video according to your own interests".

Currently, in the same order of idea that no one "Killer Application" have yet distinguish itself within the networks without that global development has been slowed down ever, no one editorial product has yet indisputably emerged into the interactive video field to constitute digital libraries databases. This field is henceforth accessible to the searchers, didacticiens, companies and editors who are willing to take the risk, and the opportunity, to exploit this new "far-west".

Bibliographical Note

After an university background on social sciences, Florent Pasquier specialised on the fields of communication sciences, multimedia technologies and educational

⁶ We recall the Alain RESNAIS's movies " SMOKING" and " NO SMOKING " representing differents possible evolutions of a story, linking them one after another, only separated by a fade to black or an off-voice : " or...".

⁷ Teaching And Learning With Interactive Media, Video and Associated Technology in School Curriculum, Report Of The Evaluation Study, September 1992 - June 1994.

sciences. Its interdisciplinary approach allows complete and global analysis, since creation of media and methods until critical observations. He obtained recompenses for its works, as the Hachette Foundation for Multimedia (1995) or the Price of Educational Media (1996).

After a few years of teaching, consulting and research engineer activities (Bouygues Group), he is now Associate Professor at the IUFM of Paris. Its last interests focuses on digital video and videoconferencing, specially for educational uses.

Bibliography

- BOURDIEU, Pierre. Sur la télévision. Liber. Raisons d'agir. 1996. 95 p.
- CHARPIN, François. "Informatique et Analyse Morphologique" *in* La pratique de l'informatique dans l'enseignement des langues anciennes, LITALA, 1995, pp. 109-125 .
- CHERFAOUI, Mourad. "Indexation et Consultation de Documents Vidéo". Thèse de Doctorat. Rennes 1. 1995.
- CLEMENT, Pierre et NDIAYE, Valdiodio. "De l'Image Vidéo à l'Image Numérisée dans des TP" Universitaires de Biologie. Les cahiers de l'ADMES N°9. 1995. pp. 115-122.
- COMPTE, Carmen. La vidéo en classe de langue. Hachette. FLE. Autoformation, 1993. 144 p.
- COSTE, Daniel. "Multimédia et Curriculum Multidimensionnel. Outils Multimédias et Apprentissage du Français Langue Étrangère". Tome 1. Cahiers de la Recherche. Université de Lille III. 1996. pp. 41-50.
- DUTEIN, Christophe. "Études et Modélisations de Décodeurs MPEG : de l'Analyse des Performances à la Synthèse d'Architecture". Thèse de Doctorat. Paris 11. 1994.
- KRISTEVA, Julia. Le langage, cet inconnu. Points. 1981. 329 p.
- LALLOUACHE, Mohamed-Tahard. "Un Poste Visage-parole Couleur : Acquisition et Traitement Automatique des Contours des Lèvres". Thèse. Grenoble. 1991.
- LE LOARER, Pierre et NORMIER, Etienne. "Techniques Linguistiques et Statistiques pour Sélectionner l'Information Pertinente". Actes du Colloque IDT'96. 1996. pp. 115-120.
- MADRANE, Nabil. "Indexation et représentation des documents vidéo". Thèse de Doctorat. ENST. Paris. 1996.
- MASSELOT, Pierre. "Les écrans du J.T". Les cahiers du CRELEF. N°26. Les discours du visible : éducation, formation, culture. 1988. pp. 6-45.
- MAZEAU, Jean-Paul. "De la Vidéo Virtuelle à la Compréhension de l'Image en Temps Réel par l'Analyse des Composantes Chromatiques : des Applications dans le Domaine de la Réalité Virtuelle". Doctorat. Paris 8. 1992.

- MOEGLIN, Pierre. "Multimédia et Education : le Démon de la Convergence. Outils Multimédias et Apprentissage du Français Langue Étrangère". Tome 1. Cahiers de la Recherche. Université de Lille III. 1996. pp. 23-40.
- PUGIBET, Véronique. "Multimédia et Civilisation". Ateliers Outils Multimédias et Stratégies d'Apprentissage du FLE. Tome 2. Université Lille 3. 1996. pp. 55-56.
- QUÉRÉ, Maryse. Systèmes-experts et Enseignement Assisté par Ordinateur (Coordination). Ophrys, AEM. 1991. 304 p.
- SEGUIER, Renaud. "Détection et localization de visages dans des séquences d'images vidéo". Thèse de Doctorat. Rennes 1. 1995.
- SIMONNOT, Brigitte. "Modélisation Multi-agents d'un Système de Recherche d'Information Multimédia à Forte Composante Vidéo". Thèse de Doctorat. Nancy 1. 1996.
- STENTON, Anthony. "L'Enseignement Assisté par Ordinateur et par Vidéo dans l'Apprentissage de l'Anglais". Doctorat. Dijon. 1992.
- STIEGLER, Bernard. Technologies de la Mémoire et de l'Imagination. 1985.
- VITALIS, André. "La Part de la Citoyenneté dans les Usages". in A. Vitalis (dir.) . Médias et Nouvelles Technologies. Apogée. Rennes. 1994. pp. 35-44.

Interaction Profiling in Digital Libraries through Learning Tools

G. Semeraro, F. Esposito, N. Fanizzi, and S. Ferilli

Dipartimento di Informatica
Università di Bari
Via Orabona, 4 I-70125 Bari
{semeraro,esposito,fanizzi,ferilli}@di.uniba.it

Abstract. We present improvements to a learning module, the *Learning Server*, to be exploited in a digital library system for supporting document management tasks as well as for providing a form of user interface adaptivity based on user classification. Indeed, our system is equipped with a web-based environment endowed with visual tools that are thought for improving the interaction of inexperienced users and for supporting experienced users in an effective accomplishment of their retrieval tasks. By logging user interaction, the Learning Server is able to suggest the most suitable interaction tools for each user.

1 Introduction

Digital libraries research focuses on the support for functions allowing access, retrieval, and organization of information, in order to make rapid decisions on what is relevant and which patterns exist among objects contained in the library. Thus, it becomes fundamental to decide which interface to assign to each user.

Among the characterizing features of CDL (Corporate Digital Library), a digital library prototype [11, 4], there is the exploitation of machine learning techniques for *document analysis*, *classification*, and *understanding* [5, 6]. Besides, the interaction environment is endowed with representations of some meta-information concerning document content, that provide users with proper cues for locating the desired data. Visual techniques are exploited, whose main advantage is the capability of shifting load from user's cognitive system to his perceptual system [10, 3].

We further improved the user-profiling module of the CDL visual environment [3], designed for providing each user with the most suitable interface that complies with his skills and knowledge about the content of a digital library. This feature is achieved by automatically classifying the user by means of machine learning techniques [10]. The Learning Server has been enriched with ultimate tools for the induction of decision trees [13].

The paper is organized as follows. Section 2 gives an overview of the architecture and the visual interaction environment of CDL. WEKA, a novel suite of machine learning and data mining tools, is presented in Section 3, while Section 4 shows experiments on interface adaptivity achieved by using the new learning component. Section 5 concludes the paper and outlines future work.

2 Overview

CDL was developed as an evolution of IDL (Intelligent Digital Library) [11, 4] in order to provide a common infrastructure supporting the creation, dissemination, manipulation, storage, integration and reuse of information. The adjective *corporate* designates a shared mechanism for searching information, updating content, controlling user access, charging users, etc., independently of the meaning and internal representation of information. The CDL project focuses on the development of effective middleware services for digital libraries, and on their interoperability across heterogeneous hardware and software platforms [2].

In CDL the typical client/server architecture of a hypertextual service on the Internet has been implemented. Since a thin-client stateful architecture [7] was adopted, it suffices a Web browser to run the application on the client-side. Indeed, there is no need of storing data locally: the DBMS runs on the server-side and its services are accessed by means of servlets. Furthermore, the architecture is characterized by a *Learning Server* that provides all the services related to document management. Besides, it is able to infer interaction profiles concerning the different types of users from data collected in *log files*. Advantages and a complete description of the architecture have been discussed in [11, 6, 3].

The main features of CDL are strictly related to the library functions. Supervised learning systems are used to overcome the problem of cheaply and effectively setting information items free of the physical medium on which they are stored (*information capture*). These learning tools are also used to automatically perform the tasks of *document classification* and *document understanding* (recognition of the logical structure of a document), that are necessary steps to index information items according to their content (*semantic indexing*). Finally, CDL provides different users with distinct interaction tools. It helps novice users to understand the content and the organization of a digital library through a suitable visual environment, and supports users that are familiar with the digital library content in easily and fast retrieving desired information items by means of an appropriate interface modality. A form of adaptivity is achieved through an automated user classification process based on machine learning techniques.

Three different roles have been identified for the interaction with CDL [4]: the *Library Administrator* runs the service and creates/deletes libraries; the *Librarians* manage specific digital libraries; and the (*Generic*) *User* requests CDL services. A user can consult the available libraries and query or browse them to retrieve documents of interest. Then, he may display/obtain the digital version of the documents found. The user interaction environment of CDL [3] is adapted according to the user class automatically recognized.

The *form-based* interface, more powerful and flexible, is appropriate for users who are already acquainted with the library structure and know about its content. Casual users often perform queries whose result is null, just because they do not have any idea of the kind of documents stored in the library. Other visual tools are better suited for novice users to grasp the nature of the information stored more intuitively and the possible patterns among the objects, so that they can make rapid decisions about what they really need and how to get it.

The *topic map* gives a global view of the semantic content of a collection of documents by exploiting a geographic metaphor [14]. A collection of items (topics or documents) is considered as a territory containing resources (the items themselves); regions, cities, and roads are used to convey an idea of the underlying semantic relationships: a region represents a set of items, and its size reflects the number of items in that region. Similarly, the distance between two cities reflects the similarity relationship between them: if two cities are close to each other, then the items are strongly related (e.g., topics are related to each other).

The *tree-based* interface provides another visual modality to both browse CDL and perform queries. The user navigates into CDL along a tree structure, starting from the root and expanding the tree step by step, so that at each node he can decide whether further exploring that path. By selecting a node, one can choose to expand it or to ask for an explanation of its meaning. Recently, this interface has been improved, giving the possibility of specializing or generalizing the currently explored search-index by means of an ontology.

3 A Learning Component Based on WEKA

WEKA (Waikato Environment for Knowledge Analysis) [13] is a system, written in Java, providing a uniform interface to different learning algorithms (*classifiers*), along with pre and post processing tools for evaluating their outcomes. Classifiers can be applied to datasets, supplied in a proper format (*arff*), made up of pre-classified instances of some class. Tools for preprocessing the data (*filters*) can be used to delete specific attributes from a dataset and to perform manual attribute selection. The learning schemes most important to our purposes are illustrated in the following.

J4.8 implements a recent improved version of C4.5 decision tree learner [9]. It outputs a decision tree in textual form (that might have been automatically pruned for achieving a better predictive accuracy) and an estimate of the tree predictive performance over both training and test datasets, generated by WEKA's evaluation module (see Figure 1). Nodes in a decision tree involve a test on a particular attribute; leaf nodes give a classification. To classify an unknown instance, a path in the tree is found starting from the root node, according to the values of the attributes tested in successive nodes, until a leaf node is reached. Classification rules represent an alternative to decision trees: the *antecedent* is a series of tests ANDed together; the *consequent* gives the class that applies to instances covered by that rule. Decision rules are generated directly off a decision tree, one for each leaf. They are usually pruned to remove redundant tests and to achieve a better predictive accuracy on further datasets. In the following we will use decision trees and rules as synonyms.

PART produces rules from partially pruned decision trees built using C4.5 heuristics. NAIVEBAYES implements the probabilistic Naive Bayesian classifier; DECISIONTABLE produces a decision table exploiting a good subset of attributes; IBK is an implementation of the k-nearest neighbors classifier (*instance-based*

```

48 pruned tree
Freq_of_Tree_Based_choices <= 0.4
| Freq_of_Topic_Map_choices <= 0.285714
| | Num_of_Parsing_Error_on_Form_Based <= 2
| | | Num_of_connections <= 1: Novice (3.0/1.0)
| | | Num_of_connections > 1
| | | | Num_of_Sort_on_Paper_of_class_Springer_of_DB_AI_in_DL <= 0
| | | | | Num_of_Query_on_DB_AI_in_DL <= 14: Teacher (31.0)
| | | | | Num_of_Query_on_DB_AI_in_DL > 14
| | | | | Num_of_Query_on_Title_of_class_Springer_of_DB_AI_in_DL <= 0: Teacher (3.0)
| | | | | Num_of_Query_on_Title_of_class_Springer_of_DB_AI_in_DL > 0: Novice (3.0)
| | | | | Num_of_Sort_on_Paper_of_class_Springer_of_DB_AI_in_DL > 0: Novice (2.0/1.0)
| | Num_of_Parsing_Error_on_Form_Based > 2: Novice (6.0/1.0)
| Freq_of_Topic_Map_choices > 0.285714
| | Num_of_Query_on_Authors_of_class_Icml_of_DB_AI_in_DL <= 1
| | | Num_of_Query_on_Paper_of_class_Icml_of_DB_AI_in_DL <= 0
| | | | Freq_of_Sort_on_Authors_of_class_Icml_of_DB_AI_in_DL <= 0.111111: Novice (70.0/6.0)
| | | | Freq_of_Sort_on_Authors_of_class_Icml_of_DB_AI_in_DL > 0.111111
| | | | | Num_of_Query_on_Abstract_of_class_Icml_of_DB_AI_in_DL <= 0: Teacher (3.0)
| | | | | Num_of_Query_on_Abstract_of_class_Icml_of_DB_AI_in_DL > 0: Novice (3.0)
| | | Num_of_Query_on_Paper_of_class_Icml_of_DB_AI_in_DL > 0: Teacher (3.0/1.0)
| | | Num_of_Query_on_Authors_of_class_Icml_of_DB_AI_in_DL > 1
| | | Freq_of_NullQuery_on_class_Springer_of_DB_AI_in_DL <= 0.076923: Teacher (3.0)
| | | Freq_of_NullQuery_on_class_Springer_of_DB_AI_in_DL > 0.076923: Novice (2.0/1.0)
Freq_of_Tree_Based_choices > 0.4
| Freq_of_Form_Based_choices <= 0.444444
| | Num_of_Help_Request_of_Form_Based <= 0
| | | Num_of_Sort_on_Abstract_of_class_Icml_of_DB_AI_in_DL <= 0
| | | | Num_of_Sort_on_Affiliation_of_class_Springer_of_DB_AI_in_DL <= 0: Expert (23.0)
| | | | Num_of_Sort_on_Affiliation_of_class_Springer_of_DB_AI_in_DL > 0
| | | | | Num_of_Query_on_Title_of_class_Icml_of_DB_AI_in_DL <= 1: Novice (2.0)
| | | | | Num_of_Query_on_Title_of_class_Icml_of_DB_AI_in_DL > 1: Expert (4.0)
| | | Num_of_Sort_on_Abstract_of_class_Icml_of_DB_AI_in_DL > 0
| | | | Num_of_Query_on_Keywords_of_class_Pami_of_DB_AI_in_DL <= 0: Expert (2.0)
| | | | Num_of_Query_on_Keywords_of_class_Pami_of_DB_AI_in_DL > 0: Novice (2.0)
| | Num_of_Help_Request_of_Form_Based > 0
| | | Num_of_Sort_on_Title_of_class_Icml_of_DB_AI_in_DL <= 0
| | | | Freq_of_Query_on_Keywords_of_class_Pami_of_DB_AI_in_DL <= 0.066667: Novice (8.0/1.0)
| | | | Freq_of_Query_on_Keywords_of_class_Pami_of_DB_AI_in_DL > 0.066667: Expert (2.0)
| | | Num_of_Sort_on_Title_of_class_Icml_of_DB_AI_in_DL > 0: Expert (2.0)
| Freq_of_Form_Based_choices > 0.444444: Teacher (2.0)

Number of Leaves : 21

Size of the tree : 41

=== Error on training data ===

Correctly Classified Instances 167 93.2961 %
Incorrectly Classified Instances 12 6.7039 %
Mean absolute error 0.0715
Root mean squared error 0.1891
Total Number of Instances 179

=== Confusion Matrix ===

 a b c <-- classified as
90 0 1 | a = Novice
3 33 0 | b = Expert
8 0 44 | c = Teacher

=== Error on test data ===

Correctly Classified Instances 12 57.1429 %
Incorrectly Classified Instances 9 42.8571 %
Mean absolute error 0.2815
Root mean squared error 0.5092
Total Number of Instances 21

=== Confusion Matrix ===

 a b c <-- classified as
5 0 5 | a = Novice
2 4 0 | b = Expert
2 0 3 | c = Teacher

```

Fig. 1. A decision tree induced by J4.8

learning); M5' implements an induction algorithm generating *model trees*, that predict the class value using a linear regression model stored at the leaf nodes.

WEKA also includes implementations of algorithms for *clustering* and for learning *association rules*. Clustering is used to group instances that seem to naturally fall together according to some similarity measure. The aim of building clusters of the training instances is to discover classes underlying the dataset in order to be able to assign further instances to them.

WEKA can be used for inducing profiles of the quality of user interaction with CDL, to build an extractor of permanent attributes (a *profile*) from logs containing session related features together with their relationships. This can be cast as a problem of supervised learning, aiming at producing rules for classifying users on the ground of their interaction. WEKA supervised learning techniques can be used for analyzing instances of interaction logs that describe features such as user daily number of connections, queries on a given search-index, etc. One of WEKA's learning schemes can be used to induce a set of rules, provided a dataset of training instances together with their classes. Then, users accessing the digital library service can be categorized on the basis of their logs, so that the most suitable interface can be chosen for them.

A different perspective regards exploiting the logs to generate more general *usage patterns* that characterize groups of users (abstracting away from individual differences) and can be used to generate general assumptions about new users. Since classes are not given *a priori*, WEKA's unsupervised learning methods could be used to extract regularities from the logs in order to find out relevant rule patterns for identifying communities of users. If the descriptions are likely to change in time, incremental approaches could be best suited.

4 The Learning Server: Interaction Profiling

CDL Learning Server can be defined as a suite of learning systems that can be exploited concurrently for performing tasks such as the mentioned document analysis, classification and understanding as well as the inference of user interaction profiles. In this section, we address this last task by means of supervised learning methods. Indeed, when developing a system accessed by several users like a service on the World-Wide Web, a fundamental problem to cope with is ensuring effective answers to the needs of various user classes. In fact, each user may have special capabilities, skills, knowledge, preferences and goals, thus he will behave differently from other users. Being hard to recognize each single user, we take into account classes of users that may interact similarly.

In the architecture of CDL, an intelligent component of the Learning Server is able to automatically assign a user to a specific class in order to improve the system usability. This component should help users at accomplishing their goal easier (through contextual helps, explanations, suitable interaction modalities etc.). As a consequence, one of the main problems concerns the definition of meaningful user classes, and the identification of the features that properly describe each of them and characterize the corresponding kind of interaction.

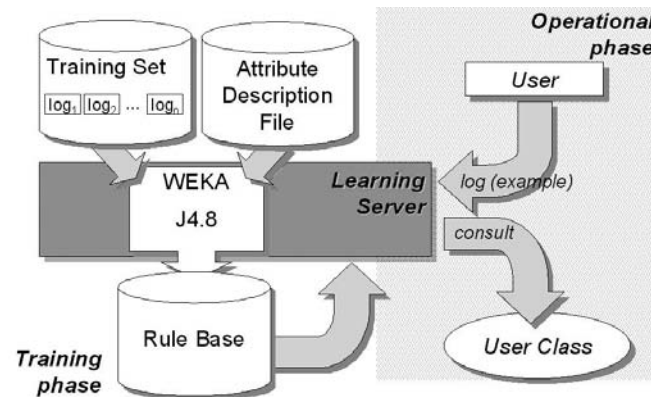


Fig. 2. Assigning user classes through the CDL Learning Server

4.1 Assigning User Classes

The main function required is to automatically assign each user to one of some predefined classes on the ground of information drawn from real interaction sessions (*interaction modeling*) [1]. Our approach relies on machine learning methods [8], since interaction modeling can be cast as a supervised learning problem by considering some user interactions with CDL as training instances for a learning system, whose goal is to induce rules for classifying CDL users (see Figure 2). Such a classification is used to associate each user class with an interface that is adequate to the user's degree of familiarity with the service. This aims at speeding up the process of understanding the organization and the content of the digital libraries of interest for an effective retrieval of the desired information.

We defined three possible classes of CDL users, namely **Novice**, **Expert** and **Teacher**. Since, during the time, the user may become familiar with the system usage, potential changes of user class should be tracked. This requires user registration and identification. Each new user has to fill in a form with personal data, then he receives an identity code - User ID - that he will use whenever he enters CDL again. Correspondingly, a log file is associated to each User ID.

By examining the log file, it is possible to extract some characteristics that are useful for recognizing the user class. We observed that most of the characteristics identified turned out to be application dependent, while only few showed to be system dependent. For instance, relevant characteristics are those concerning the way users exploit the capabilities of CDL search engine, such as date and time of session beginning, document class or search indexes chosen, criterion for sorting the search results, number of documents obtained as results of the search, types of errors made during the interaction.

Log files are then exploited to train a learning system in order to induce a decision tree and a set of rules that are used by the system to autonomously classify users. Figure 3 reports a portion of an *arff* file, containing the attributes and the corresponding values coming from the users log files used for our purposes.

[illegible]

Fig. 3. A portion of the *arff* file used for user classification

4.2 Experimental Results

Following the preliminary experiment described in [3], both C4.5 [9] and the WEKA classifier J4.8 [13] have been customized in order to infer the classification rules in a batch way. Another experimentation has been set up with ITI 2.0 [12], a system which supports the incremental construction of decision trees by revising the current tree in response to each newly observed training instance. ITI can operate in two different incremental ways. In the normal operation mode, it first updates the frequency counts associated to each node of the tree as soon as a new instance is received, then it restructures the decision tree according to the updated frequency counts. In the Error Correction (EC) mode, the frequency counts are updated only in case of misclassification of a new instance. The main difference is that the normal operation mode guarantees the building of the same decision tree independently of the order in which the examples are presented, while the error-correction mode does not.

The new experiment on classification of CDL users involved 200 instances of interaction, 90% of which were drawn as training set, while the rest was used as test set. Each log file was used to draw 128 attribute values that describe user interactions and it was associated to one of the three classes mentioned. Then, the classifiers have been run in parallel on the same training and test sets for a comparison of the outcomes. The experiment has been replicated ten times by growing the decision trees on the training set, then pruning them. The performance of both unpruned and pruned trees has been evaluated by using them for classifying instances in the test set.

Table 1 reports the results of the experiment, in terms of predictive accuracy on the test set. A first comparison between the results obtained by C4.5 and J4.8 reveals that the latter outperforms the former in half of the runs, while the opposite happens in only 3 cases (2 runs are even), both for the pruned and for the unpruned case. In fact, such a predominance is confirmed by the mean predictive accuracy only for the pruned case, probably due to the very good performance reached in run #1 by C4.5 unpruned. On the other hand, comparing such systems with the incremental ones results in a clear success of the latter, whose predictive accuracy is 2.4% better in the worst case (ITI unpruned vs. C4.5 unpruned), reaching a 4.4% improvement in the best case (ITI EC pruned vs. C4.5 pruned). In particular, as regards ITI, the error correction modality outperforms by 0.4% the normal one when no pruning is allowed, while there is no clear difference between the two in the pruned case.

After the training phase, whenever any user accesses CDL through a client, the log file generated during the interaction session is exploited to provide a new example that the Learning Server will classify on the ground of the inferred rules. The way in which rules are consulted by the Learning Server (and the existence of the classification rules itself) is completely transparent to the user. On the ground of the classification, the Learning Server selects a distinct type of interface, regarded as the proper one for that user class. Specifically, the Learning Server suggests to prompt any user recognized as a member of the class *Novice*

Table 1. Outcomes of the experimentation (Predictive Accuracy)

	0	1	2	3	4	5	6	7	8	9	mean
C4.5 unpruned	47.6	95.2	66.7	71.4	66.7	61.9	66.7	71.4	47.4	66.7	66.2
J4.8 unpruned	66.7	66.7	66.7	81.0	71.4	47.6	66.7	81.0	36.8	76.2	66.1
ITI unpruned	85.7	71.4	66.7	66.7	66.7	61.9	66.7	66.7	52.6	76.2	68.1
ITI EC unpruned	71.4	81.0	61.9	85.7	71.4	47.6	66.7	71.4	47.4	81.0	68.6
C4.5 pruned	47.6	95.2	66.7	76.2	66.7	61.9	66.7	76.2	42.1	76.2	67.6
J4.8 pruned	57.1	85.7	71.4	81.0	71.4	52.4	66.7	81.0	36.8	76.2	68.0
ITI pruned	85.7	76.2	76.2	76.2	71.4	57.1	57.1	81.0	52.6	85.7	71.9
ITI EC pruned	76.2	90.5	42.9	81.0	76.2	71.4	52.4	85.7	57.9	85.7	72.0

with the topic map interface, the tree-based interface is proposed to a user in the class **Expert** and **Teacher** users have the form-based interface as a default.

The main idea underlying the mapping between user classes and types of interfaces is that users being unfamiliar with the system need an environment that allows them to preliminarily understand the content of a digital library. More skilled users are supposed to be already acquainted with both the organization and the content of the digital library, thus they need tools that allow to speed up the retrieval process. Independently of the decision made by the system classifier, any user is free to switch to another interface at any moment.

The system showed the ability to track user class evolution. Indeed, after a number of interactions, CDL may propose a different interface to the same user, as soon as more information are collected in his log.

5 Conclusions and Future Work

We presented a component of the Learning Server exploited in our digital library CDL for performing a form of user profiling. Indeed, CDL is equipped with a web-based visual environment, offering different solutions to differently skilled users. An effective user classification as that provided by a learning component may be fundamental for providing a form of interface adaptivity. Both systems that operate in a batch way and incremental ones have been tested on such a task.

We intend further study the use of incremental learning schemes, that avoid the drawback of starting over the learning process from scratch each time new log examples become available, and hence are more suitable for the nature of Digital Library applications. Besides, being the WEKA based on Java technology, it seems suitable for an exploitation for online learning. Furthermore, also unsupervised learning methods could be exploited to cluster instances together to form *usage patterns*, thus inferring also the number of classes directly from the available data. Finally, we plan to apply learning tools to the inference of the ontologies mentioned in Section 2. This would make the search through the tree-based interface more knowledge-intensive.

Acknowledgments. The authors would like to thank Francesca Lioce who performed the experiments with the CDL Learning Server.

This work was partially funded by the IST Project COGITO 1999-13347.

References

- [1] D. Banyon and D. Murray. Applying user modeling to human-computer interaction design. *Artificial Intelligence Review*, 7:199–225, 1993.
- [2] P. A. Bernstein. Middleware: A model for distributed system services. *Communications of the ACM*, 39(2):86–98, 1996.
- [3] M.F. Costabile, F. Esposito, G. Semeraro, and N. Fanizzi. An adaptive visual environment for digital libraries. *International Journal of Digital Libraries*, 2(2+3):124–143, 1999.
- [4] M.F. Costabile, F. Esposito, G. Semeraro, N. Fanizzi, and S. Ferilli. Interacting with IDL: The adaptive visual interface. In *Research and Advanced Technology for Digital Libraries. Proceedings of the 2nd European Conference - ECDL98*, volume 1513 of *Lecture Notes in Computer Science*, pages 515–534. Springer, 1998.
- [5] F. Esposito, D. Malerba, and G. Semeraro. Multistrategy learning for document recognition. *Applied Artificial Intelligence*, 8(1):33–84, 1994.
- [6] F. Esposito, D. Malerba, G. Semeraro, N. Fanizzi, and S. Ferilli. Adding machine learning and knowledge intensive techniques to a digital library service. *International Journal of Digital Libraries*, 2(1):3–19, 1998.
- [7] M.C. McChesney. Banking in cyberspace: An investment in itself. *IEEE Spectrum*, pages 54–59, February 1997.
- [8] V.S. Moustakis and J. Herrmann. Where do machine learning and human-computer interaction meet? *Applied Artificial Intelligence*, 11:595–609, 1997.
- [9] J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.
- [10] G. Semeraro, M. F. Costabile, F. Esposito, N. Fanizzi, and S. Ferilli. A learning server for inducing user classification rules in a digital library service. In Z.W. Raś and A. Skowron, editors, *Proceedings of the 11th International Symposium on Methodologies for Intelligent Systems*, volume 1609 of *Lecture Notes on Artificial Intelligence*, pages 208–216. Springer, 1999.
- [11] G. Semeraro, F. Esposito, D. Malerba, N. Fanizzi, and S. Ferilli. Machine learning + on-line libraries = IDL. In C. Peters and C. Thanos, editors, *Research and Advanced Technology for Digital Libraries. First European Conference - ECDL97*, volume 1324 of *Lecture Notes in Computer Science*, pages 195–214. Springer, 1997.
- [12] P.E. Utgoff, N. Berkman, and J.A. Clouse. Decision tree induction based on efficient tree restructuring. *Machine Learning*, 29:5–44, 1997.
- [13] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.
- [14] M. Zizi and M. Beaudouin-Lafon. Hypermedia exploration with interactive dynamic maps. *International Journal on Human-Computer Studies*, 43:441–464, 1995.

DEBORA: Developing an Interface to Support Collaboration in a Digital Library

David M. Nichols¹, Duncan Pemberton², Salah Dalhoumi³,
Omar Larouk³, Claire Belisle⁴, and Michael B. Twidale⁵

¹ Department of Computer Science, University of Waikato, Hamilton, New Zealand
daven@cs.waikato.ac.nz

² Computing Department, Lancaster University, LA1 4YR, UK
pemberto@comp.lancs.ac.uk

³ École Nationale Supérieure des Sciences de L'Information et des Bibliothèques,
69623 Villeurbanne Cedex, France
{dalhoumi, larouk}@enssib.fr

⁴ LIRE, Unité Mixte de Recherche 5611, CNRS-Université Lyon 2, France
claire.belisle@ish-lyon.cnrs.fr

⁵ Graduate School of Library and Information Science,
University of Illinois at Urbana-Champaign, IL 61820, USA
twidale@uiuc.edu

Abstract. Interfaces to library systems have largely failed to represent the inherently collaborative nature of information work. This paper describes how collaborative functionality is being implemented as part of the DEBORA project to provide access to digitised Renaissance documents. Work practices of users of Renaissance documents are described and the collaborative features of the client software are outlined. Functionalities discussed include annotation, the creation of virtual books and the inclusion of user-supplied metadata.

1 Introduction

This paper describes the development of collaborative functionality for users of digital libraries in the context of the EU *Telematics for Libraries* project DEBORA (Digital Access to Books of the Renaissance).

The aim of the DEBORA project is to make Renaissance books more generally available as digital resources and to examine the potential for novel collaborative functionality. The collection being created within DEBORA consists of digitised images of books from libraries in Lyon, Rome and Coimbra.

The first part of the paper outlines the nature of collaboration in digital libraries. Section 3 describes evidence gained from real life users of Renaissance materials. Section 4 describes the implementation of collaborative functionality in the DEBORA client followed by some initial user studies and a conclusion.

2 Collaboration in Digital Libraries

Digital Libraries offer new opportunities for collaboration and communication that were unfeasible in traditional libraries. Interfaces to information systems (including databases, library catalogues and information retrieval applications) have largely reflected single user stereotypes [27]. That is, the activities of other users have had almost no impact on the experience of any one user.

The technology associated with Digital Libraries allows us to consider new ways of working with library materials [13]. Specifically, for users to work in groups, rather than individually, and for users to add to collections, rather than simply reading.

2.1 Overview of Collaborative Work for Digital Libraries

Digital Libraries, in comparison with print-based libraries, more easily support the modification of their contents. Several researchers (e.g. [27, 8, 9, 10]) have recognised the potential for users, rather than librarians, to contribute to the development of a collection through user-supplied data (USD) [10]. Such USD can come in many different forms, although it can be split into two main groups: data automatically collected from users' activities and data explicitly generated by users. Implicit additions to a collection include: search term suggestion [10], ratings [19] and 'read-wear' [7]. There have been many proposals for explicit USD: annotation [30, 14, 2], keyword addition [4, 10], evaluative commentary [9, 10], hypertext links [3, 9, 8, 18, 15, 24, 23], ratings [23, 2], error correction [26] – for a review see [27].

The common thread amongst these ideas for collaboration is that one users' actions can be shared with other users *within the system* [27, 13]. In a paper-based library such sharing is more circuitous – via the publishing of cross-cited works that eventually physically arrive in libraries. The belief amongst researchers is that such collaboration will be more productive for the users [13]: e.g., by enhancing retrieval effectiveness through community rating of resources [2]. Although many forms of USD have been suggested the most common example is probably *annotation*.

Annotations. Annotation has been frequently proposed as a technique for users to add content (and so share ideas) within information systems – for a review see [22]. However, as Wilensky recently notes: “despite its evident usefulness, digital annotation capabilities are not very widespread” [30]. There are several annotation systems in different contexts [22] but the ‘broad territory’ of annotation [15], from free text to metadata, has not been conducive to the development of an accepted standard for annotations [22]. Marshall [15] characterises annotations along seven dimensions: formal/informal, explicit/tacit, permanent/transient, public/private etc. Wilensky [30] suggests four system requirements; annotations should be:

- Able to be placed *in situ*
- Expressive, extensible and composable

- Format and platform independent
- Free of permission and registration requirements

That annotations should be placed (and viewed) *in situ* accords with evidence from real world studies [14, 15]. The second and third requirements follow from the variability of users, usage contexts and documents; although they also reflect the generality of the *multi-valent document* approach [21] and the desire to support spontaneous ubiquitous collaboration [30]. The requirement that annotations should not be dependent on prior registration, illustrates a particular perspective on the usage of annotations. If we wish to collaboratively construct cataloguing information [4] (the formal end of Marshall's dimension), then user registration may be necessary to maintain the metadata authority.

The different interpretations of 'annotation' cover many of the forms of explicit USD. Consequently they share attributes with other novel collaborative functionalities: presenting significant added complexity to designers, a lack of accepted standards and their adoption would entail significant changes to users' work practices.

2.2 Designing for New Ways of Working

By studying existing practice and comparing it with evolving practice in other disciplines (especially in the sciences) that have enjoyed better computational resources over a longer period of time, we can explore the design space and create systems that 'add value' to existing work activities. Our design challenge is to support new ways of working by adding collaborative features, but they must be incorporated in a thoughtful manner. A digital library that allowed users to contribute and share information around collection items would be a radical change for many users – especially in a domain such as Renaissance texts. Humanities scholars mostly do not perceive themselves as working collaboratively [20]. Thus just asking them about their collaborative work is insufficient; that work must also be observed in order to see the substantial 'invisible collaboration'. The system must allow different kinds of use, from the currently conventional solitary forms of work, to supporting more effectively existing kinds of collaborative work, to supporting kinds of collaboration wholly new to this group. To be acceptable, the system needs to support graceful (and if necessary, slow) transitions in use along that scale.

3 Work Practice and Design Implications

Renaissance books represent a turning point not only because the advent of printing brought about a gradual re-organisation and standardisation of the textual material and its presentation [16], but also because the political, social and religious upheavals which characterise the century profoundly influenced the use of books.

Due to their rarity and fragility, the availability of Renaissance texts has generally been limited to acknowledged scholars. But even these researchers can have access

difficulties. The research work done on such book corpuses is basically comparing different versions, identifying and tracing originality and influences [25, 29].

At the same time, there is an increasing number of requests for access to 16th century material coming from a variety of users including: educators and their students, linguists, book historians, social and cultural historians ('histoire des mentalities'), specialists in literary studies, illustrators, wardrobe designers etc. [5].

This section summarises a usage study undertaken through observations, interviews and questionnaires with users familiar with Renaissance or other old books. The main findings are taken from the results of a survey through a written questionnaire (83 questions) answered by 62 scholars working on old or 16th century books.

3.1 Observations of the Use of Renaissance Documents

Working with Renaissance books implies, for specialists, accessing a specific copy, for there are almost no two identical copies that have survived (even copies of the same text from the same publisher). Each copy will reveal unique information on where, when and how it was printed, through explicit information in the book or through the book's material composition.

Two groups of users of Renaissance books can be identified: book specialists (who may eventually require access to the physical copy) and those interested in the content. This second (larger) group, including users of the existing digitised networked material [29], has working habits that can be summarised in four main areas:

- They need to be able to identify the specific copy of the book they are accessing and very carefully read the text or study the illustrations [17].
- While doing this, they take notes, either handwritten or on their computers, if these are allowed in the library (and if using local electrical outlets is permitted).
- Scholars find themselves alone, or almost, in their speciality: for example, the study of the 16th century dialect of Lyon in the writings of Paradon, a regional bishop. Consequently, they will not tend to exchange information - often because there is no-one to share it with. This may also be related to a university tradition of individual evaluation based on personal, and not collaborative, publications.
- Each scholar usually maintains a system of indexed cards where she stores all the information patiently gathered from many trips to libraries and archives; together with any personal notes. Some scholars now use computers for this task.

3.2 Current Collaborative Practices

There is an increasing recognition of the collaborative aspects of many forms of work [12, 27, 28]. We expect that the benefits potentially available from networked interaction will increase collaboration and modify the manner in which users work. We found that almost 80% of the scholars use some form of Web access to find and exchange information. Although they rarely have the occasion to collaborate specifically in their

research, email is used widely. These collaborations often involve passing on, to a colleague, information found by chance (serendipitous finds).

Collaboration in real time (synchronous conferencing) is not seen as an essential function. Answers focus on a more restricted sense of collaboration; finding out how the work of colleagues is going and asking information on certain aspects of their own work. A majority of scholars are willing to share their notes with colleagues and, of those who intend to collaborate, most express a desire for these tools to be integrated into the interface of the access software.

3.3 Metadata

Image collections, such as DEBORA, often lack detailed metadata. The librarians in the project are supplying typical book-level MARC metadata. Beneath this is the internal structure of a book, specific to 16th century texts: the location of indices, prefaces etc. However for effective retrieval using conventional searching the detail of individual pages (illustrations, decorative elements etc.) is needed to allocate indexing terms. Although basic structuring, such as differentiating illustrations from normal text, can be achieved with image analysis tools, most detail must be contributed by content specialists. As for most image collections, this amount of effort is infeasible. It is at this level, of detailed page-specific metadata, that collaborative contributions by the users of image collections could prove most valuable.

3.4 Design

Based on our findings, we focus on annotation mechanisms to explore collaboration. Annotation is already a part of existing solitary scholarly practice, and so potential users are most likely to be willing to learn to use the system in order to obtain the benefits of familiar work practices. This is our ‘Trojan Horse’ for studying new collaborative features based on those same annotations. Firstly, we explore the kinds of annotations that scholars find useful, through iterative prototyping.

4 The DEBORA Client

DEBORA is based around a client-server architecture with two distinct types of server. A Z39.50 based server is used for the storage of ‘official’ metadata – including the location of the catalogued images. In common with many other annotation systems a separate server is used to store and retrieve user annotations [22]. The client has two main functions: to provide access to the images of the collection and to support the collaborative functionality of the system.

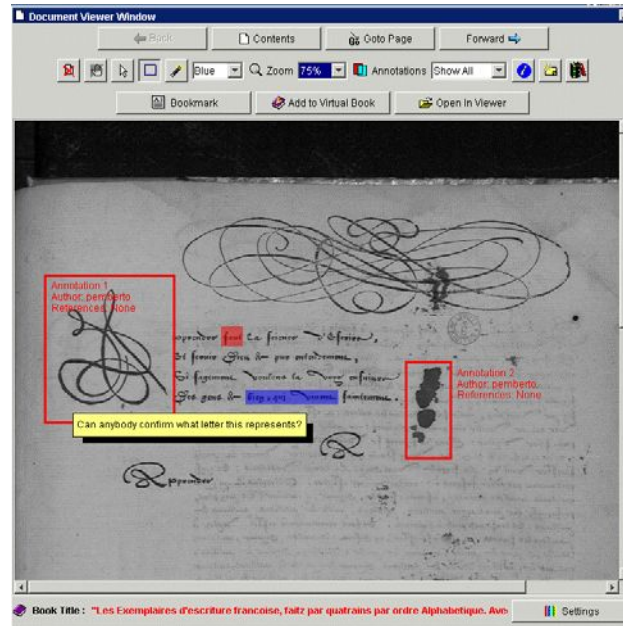


Fig. 1. The DEBORA client interface showing annotation and highlighting

4.1 Client Interface

The main window of the DEBORA client contains several image viewing tools (magnification, contrast etc). Fig. 1 shows an annotation attached to a rectangular area of a document. Annotations are currently free-text, as opposed to structured thesaurus terms [2]. The personal-public dimension of an annotation (author specified) is shown by colour and can be used to view a subset of all of the annotations.

The client also provides facilities for highlighting areas of an image in a variety of colours – typical of paper annotations [14, 15]. User-definable workspaces are provided to allow users with similar interests to structure their collaboration activities. Alternatively a user can restrict their additions to be personal and private.

Any set of annotations can be chained together to provide trails, or paths, [3, 24, 23]; following a path may involve moving to any part of any book in the collection. These hyperlinked annotations and associated images can be gathered together by a user to create a *virtual book*. We expect that this aggregation will help to reduce the adverse navigational effects of traversing a trail that spans many collection items.

Fig. 2 shows the virtual books in the lower left corner, beneath ‘real’ books from the collection. A user creates a virtual book by selecting elements (such as pages) from existing resources and arranging them in ‘virtual chapters’. This composite virtual document [18] can then become part of the collection for other users. A typical

educational scenario would be a professor tracing the historical development of an artistic style and collating examples into a virtual book for her students [24].

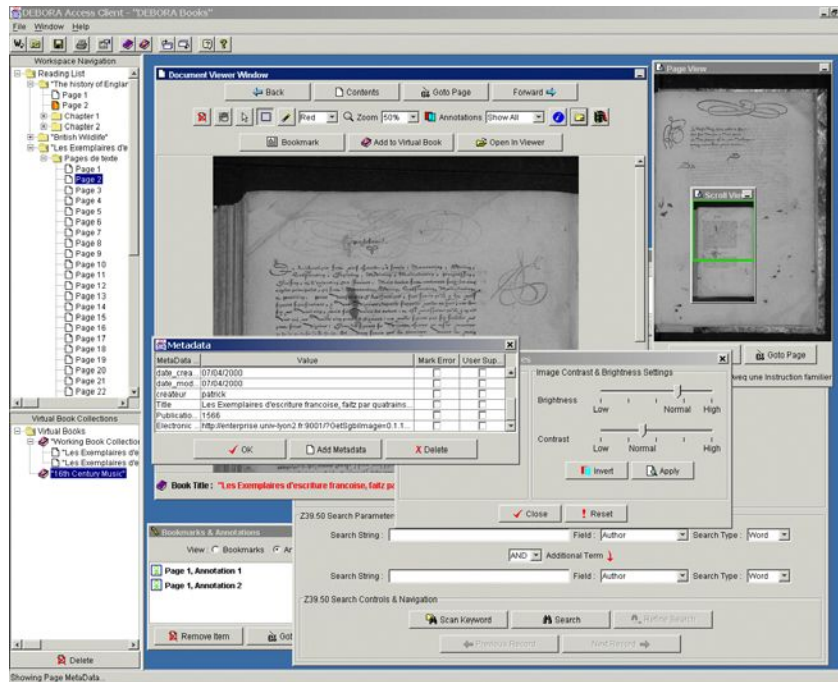


Fig. 2. The DEBORA interface showing the creation of a *virtual book* (bottom left)

The client currently displays virtual books separately from the main collection – however, if we are to take the promise of the digital library seriously then these books should be seen as *part* of the main collection. Extending virtual books to include items from other collections, and the issues of generating metadata for such documents, implies consideration of an extensible notation such as XML [22].

4.2 Collaboration and Metadata

In addition to shared annotations and virtual books there are at least three other methods for USD to enhance the metadata of the DEBORA collection.

Error Correction. Most database users have no way to record the presence of errors they detect in item descriptions. The client currently supports one-click boolean ‘error-present’ actions and allows users to suggest alternative descriptions. With a population of active users any data quality effort could prioritise those items with the most reported errors [26], or *error-wear* by analogy with *edit-wear* [7].

User-Supplied Metadata. In addition to correcting existing metadata, the client can accept new keywords from users. At present these terms are not easily integrated with metadata on the main DEBORA server but this is a small technical problem. A simple interface is required to allow human authorisation of additional metadata.

Re-purposing Annotations. The annotations are stored separately and so can be easily searched independently of the collection metadata. Annotation databases are potentially valuable sources of text [11, 22]; particularly for image databases. Golovchinsky et al. [6] use the text of freeform annotations as a source of query terms. Conversely, we can mine user annotations for indexing terms; as ‘multimedia annotations...are simply meta-data associated with multimedia content’ [1].

A major difference between conventional image annotation and this approach is the purpose of the annotation: collaborative annotations are not *intended* to describe the images. Although re-purposed annotations will generate index terms of lower quality than an expert cataloguer, it may well be better than their complete absence.

When user supplied data is used in conjunction with ‘official’ descriptions then search tools need to be aware of the differences in term authority [4]. One approach would be to attach less weight to USD terms in query matching. If the USD is accepted by collection maintainers then this increased trust could be reflected by increasing the weights of other USD terms from the same user [26].

5 Initial User Studies with the Client Interface

The client has been scrutinised by several 16th century specialists and has yielded information on its functionality. Virtual documents, where each user creates her own document by selecting pieces of one or several books, are considered essential. A virtual book is perceived as answering important and specific user needs, and is seen as close to ways of working with physical documents. Annotation is appreciated as a fundamental functionality allowing collaboration. The three levels provided, private, group and public, all address specific user concerns. Users expressed concern about managing large numbers of annotations and recognising different annotation types.

The facility to have two pages side-by-side, or two illustrations taken from two different copies of the same book, eases comparison. This is particularly appreciated by specialists who need to authenticate a dubious copy or restore a copy.

The zoom function is seen as very useful. It allows a rapid assessment of the content of the different pages presented and is also an effective replacement for leafing through a book. The possibility of rapidly going back or forward facilitates memorising and global comprehension of the document. Users working on illustrations or typography suggested an option showing the actual size of a page element. These specialists also appreciated the highlighting facility that augments the contrast and sharpness necessary in their tasks. Some expressed fear that digitising books would result in a clean image that would not be sufficiently faithful to the original.

6 Conclusions

The paper considers how collaborative features can be added to a digital library. One pertinent problem is that the intended users do not perceive themselves as working collaboratively (even when they do), and so are unlikely to see the benefits of collaborative features. We are exploring the provision of annotation features as a mechanism to support a graceful transition from solitary use to collaborative work.

Future work will involve a continuation of the refinement of the basic annotation features for conventional use, along with an exploration of the additional collaborative use of these same annotations (e.g. as a source of index terms).

Conventionally, rare texts and their annotations were separate, in DEBORA we have brought them together. It will be important to integrate other aspects of work: the organisation of annotations, the creation of virtual books and texts written by users of the digital library. Such a richer environment will not only promote easier switching between its components, but it will itself afford new forms of collaboration, and in a networked digital library mean that the user's office then becomes as mobile as the Renaissance books themselves have become.

Acknowledgements

The DEBORA project is funded under the EU *Telematics for Libraries* programme.

References

1. Barger, D.M., Gupta, A., Grudin, J., Sanocki, E., Li, F.: Asynchronous Collaboration Around Multimedia and its Application to On-Demand Training. Microsoft Research Technical Report 99-66. (1999)
2. Bouthors, V., Dedieu, O.: Pharos, a Collaborative Infrastructure for Web Knowledge Sharing. In Proceedings of ECDL'99. Springer Verlag, Berlin (1999) 215-233
3. Bush, V.: As We May Think. *The Atlantic Monthly*. 6 (1945) 101-108
4. Chandrinos, K., Immerkær, J., Dörr, M., Trahanias, P.: A Visual Technique for Annotating Large-Volume Multi-Media Databases – A Tool for Adding Semantic Value to Improve Information Filtering. In Proceedings of the 5th DELOS Workshop: Filtering and Collaborative Filtering. ERCIM, Le Chesnay, France (1998) 125-129
5. Chartier, R.: *Les Usages de L'Imprimé*. Fayard, Paris (1986)
6. Golovchinsky, G., Price, M.N., Schilit, B.N.: From Reading to Retrieval: Freeform Ink Annotations as Queries. In Proceedings of SIGIR'99. ACM, New York (1999) 19-25
7. Hill, W.C., Hollan, J.D.: History-Enriched Digital Objects: Prototypes and Policy Issues. *The Information Society*. 10 (1994) 139-145
8. Kantor, P.B.: The Adaptive Library Network Interface: A Historical Overview and Interim Report. *Library Hi Tech*. 11 (1993) 81-92

9. King, G., Kung, H.T., Grosz, B., Verba, S., Flecker, D., Kahin, B.: The Harvard Self-Enriching Library Facilities (SELF) Project. In *Proceedings of Digital Libraries '94 (DL'94)*. Texas A&M University (1994) 134-138
10. Koenig, M.E.D.: Linking Library Users: a Culture Change in Librarianship. *American Libraries*. 21 (1990) 844-849
11. Lawton, D.T., Smith, I.E.: The Knowledge Weasel Hypermedia Annotation System. In *Proceedings of the Fifth ACM Conference on Hypertext*. ACM, New York (1993) 106-117
12. Leplat, J.: *L'Analyse Psychologique du Travail: Quelques Jalons Historiques*. *Le Travail Humain*. 56 (1993) 115-131.
13. Levy, D., Marshall, C.C.: Going Digital: A Look at the Assumptions Underlying Digital Libraries. *Communications of the ACM*. 38 (1995) 77-84
14. Marshall, C.C.: Annotation: from Paper Books to the Digital Library. In *Proceedings of the 2nd International Conference on Digital Libraries*. ACM, New York (1997) 131-140
15. Marshall, C.C.: Toward an Ecology of Hypertext Annotation. In *Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia*. ACM, New York (1998) 40-49
16. Martin, H.: *Histoire et Pouvoirs de L'Écrit*, Bibliothèque de L'Évolution de L'Humanité. Albin Michel, Paris. (1996)
17. Martin, H.: *La Naissance du Livre Moderne, XIVe – XVIIe Siècles*. Electre, Paris (2000)
18. Myaeng, S.H., Lee, M., Kang, J.: Virtual Documents: a New Architecture for Knowledge Management in Digital Libraries. In *Proceedings of the Second Asian Digital Libraries Conference*. National Taiwan University, Taipei (1999) 171-181
19. Nichols, D.M.: Implicit Ratings and Filtering. In *Proceedings of the 5th DELOS Workshop: Filtering and Collaborative Filtering*. ERCIM, Le Chesnay, France (1998) 31-36
20. Palmer, C.L., Neumann, L.J.: Exploration and Translation: the Research Work of Interdisciplinary Humanities Scholars. *Library Quarterly* (to appear)
21. Phelps, T.A., Wilensky, R.: Multivalent Annotations. In *Proceedings of ECDL'97*. Springer Verlag, Berlin (1997) 287-303
22. Ovsiannikov, I.A., Arbib, M.A., McNeil, T.H.: Annotation Technology. *International Journal of Human-Computer Studies*. 50 (1999) 329-362
23. Röscheisen, M., Morgensen, C., Winograd, T.: Beyond Browsing: Shared Comments, SOAPs, Trails, and On-line Communities. *Computer Networks and ISDN Systems*. 27 (1995) 739-749
24. Shipman, F.M., Furuta, R., Brenner, D., Chung, C., Hsieh, H.: Guided Paths Through Web-Based Collections: Design, Experiences, and Adaptations. *Journal of the American Society for Information Science*. 51 (2000) 260-272
25. Sordet, Y.: Repérage et Navigation dans L'Espace du Livre Ancien. *Communication Présentée au 1er Forum de L'Édition et de la Documentation Spécialisé*. (1997)
26. Twidale, M.B., Marty, P.F.: An Investigation of Data Quality and Collaboration. Technical Report UIUCLIS--1999/9+CSCW, University of Illinois at Urbana-Champaign. (1999)
27. Twidale, M.B., Nichols, D.M.: Computer Supported Cooperative Work in Information Search and Retrieval. In Williams, M.E. (ed.): *Annual Review of Information Science and Technology (ARIST)*, Vol. 33. Information Today, Inc., Medford, NJ (1998) 259-319
28. Twidale, M.B., Nichols, D.M., Paice, C.D.: Browsing is a Collaborative Process. *Information Processing & Management* 33 (1997) 761-783
29. de Ventabert, G.: Représentation et Exploitation Électroniques de Documents Anciens (Textes et Images). *Document Numérique*. 3 (1999) 57-73
30. Wilensky, R.: Digital Library Resources as a Basis for Collaborative Work. *Journal of the American Society for Information Science*. 51 (2000) 228-245

Children as Design Partners and Testers for a Children's Digital Library

Yin Leng Theng, Norliza Mohd Nasir, Harold Thimbleby, George Buchanan,
Matt Jones, *David Bainbridge, and **Noel Cassidy

School of Computing Science, Middlesex University, U.K.
{y.theng, n.mohd-nasir, h.thimbleby, g.buchanan and m.jones}@mdx.ac.uk

*Department of Computer Science, University of Waikato, Hamilton,
New Zealand, davidb@cs.waikato.ac.nz; **St.Albans School, U.K.,
NJCASSIDY@st-albans.herts.sch.uk

Abstract. Most of today's digital libraries (DLs) are not designed for children. To produce usable and useful DLs, designers need to ensure that good design features are incorporated, taking into consideration users' needs. We describe our experience working with children as *design partners* and *testers* in building a children's DL of stories and poems for 11-14 year olds, using a concrete example to demonstrate our design philosophy and research approach. The study provides insights on useful design features children's DLs should have, and their importance to children. The initial work we have done highlights issues and provides a basis for the building of usable and useful digital libraries for children.

1 Digital Library Technology and Children

Historically, the creation of public libraries in the early 1900s was to support education by providing free access to information. With the introduction of the World Wide Web (also known as the Web) in the mid 1990s, we were presented with a powerful alternative to traditional libraries. The Web makes it "very easy" for anyone to publish and retrieve information. To cope with what is now becoming an information glut on the Web, subject-based digital libraries (DLs) are beginning to emerge on the Web. However, most of today's DLs are not designed for children [1].

Children have their own likes, dislikes, curiosities and needs that are different from adults. Designers of this new DL technology, therefore, should regard them as an entirely different user population with their culture, norms and complexities.

Using a concrete example to demonstrate our design philosophy and research approach, we describe our work in the design of a DL of stories and poems for children aged 11-14 years old. This work was carried out as part of a project funded by the UK's Engineering and Physical Sciences Research Council (EPSRC). Our DL illustrates two objectives:

- To engage children as *design partners* to ensure that the DL is designed with and for children (see Sections 2 and 3); and
- To engage children as *testers* to evaluate three different interface designs to the DL and to rate in order of importance design features crucial to the success of the DL (see Section 4).

2 Participatory Design — Engaging Children as Design Partners

In recent years, we see more research focusing on understanding the role of children in the technology design process as users, testers, informants and design partners [e.g. 2, 4, 6, 7, *etc.*]. From the start, we wanted our project to be a thoroughly collaborative endeavour as we wanted to design a DL with and for children. Hence, we invited a class of twenty-three 12-year old boys and their English teacher at a secondary school to be our design partners. These children were selected because they were competent web users and would be able to give more informed comments on the efficiency and effectiveness of DLs, compared to say, novice users. (Also, one of this paper's authors is a Governor of the school).

Two sessions were conducted during a one-hour-ten-minute English lesson between November and December 1999. The objectives of the first two sessions were to carry out *participatory design, engaging children as design partners* to:

- investigate useful design features for a children's DL of stories and poems; and
- propose and implement basic design features for the children's DL that will encourage active engagement and participation, taking into account children's needs.

Protocol of first session

We wanted to create the trust required to have a truly participatory design process and that takes time and effort [7].

In the first session, the children were introduced to the concept of DLs. We explained our research project and objectives. The students felt important at being invited to be design partners. We wanted the children to work in teams to brainstorm design ideas for the DL environment with as little input from us as possible. Teams of five or six children were formed. The teams had two weeks to think about the design features. To help them with the assignment, they were provided with URLs to browse sample DLs and children's websites. The children were to submit their design ideas before the second session.

Protocol of second session

The second session took place three weeks after the first session. Each team presented the design ideas proposed for the DL environment, which they had discussed in their groups after the first session.

Results and Analyses

What the children wanted in the DL was entertainment and fun. One suggestion was that the library could be divided into different areas: (a) study area to allow reading or browsing; (b) librarian area to make enquiries; (c) games area to play games; (d) dictionary area to search for meanings of words; and (e) actual library to search for books. Another suggestion was to provide different forms of help such as having a librarian at the "front desk", a search box and a random book selector. The children also suggested that the DL environment to include a display of the top ten stories/poems, information about authors and a message board to post and discuss ideas.

Table 1 summarises the ideas submitted by the four teams (see <http://www.cs.mdx.ac.uk/dl/St.Albans/> for work done by the children). At the end of the second session, a list of design features the children liked to be included in the collaborative writing environment was formed (see Table 1). These features were implemented in the three different interfaces to our DL prototype, described in more detail in Section 4.

Table 1. Ideas proposed by the teams, and list of design features implemented in the prototype.

Team	Ideas proposed by the teams	List of design features implemented in our DL prototype (see Section 4)
A	- Bookshelf organised in different sections (e.g. science fiction, fantasy, etc.) - Facility to write and submit stories - List of top 10 books/authors - Good index feature - Provide interesting and fun activities such as a picture of man falling over, books falling off the bookshelves, etc. - Provide a random book selector with the relevant book opened to be read	Design DL to be: - like "traditional" library - more game-like - efficient in searching for relevant materials Offer children opportunities to submit to the DL.
B	- Facility to write/submit story - List of books with reviews and comments from readers - Search for relevant materials giving either information on "title" or "author" - Facility to contact the author via email to give feedback	Give recognition of good stories submitted by listing the top 10 books/authors. Offer children with fun features to search for relevant books, etc.
C	- Provide games - List of top 10 books, authors - Links to other relevant sites and writing aids such as dictionary, thesaurus - Facility to chat with other users	Provide opportunities to chat with and to get feedback from other readers.
D	- Create a virtual library with study area, librarian area, dictionary area, search area and games area - Select a weekly winning book - Provide audio feature for the story to be read - Provide a random book selector	

3 Design Choices

Using inputs from the children as well as drawing upon design guidelines highlighted by other researchers working with children [e.g., 2, 4, 5, 6, 7, *etc.*], we describe our DL environment with distinctive features to provide opportunity: (1) to create and submit stories/poems to the DL, collaborate with other users in writing stories/poems and actively engaging users in the use of the DL; and (2) for flexibility and manipulation of search results.

3.1 Provide Opportunity for Creation, Collaboration, and Engagement

Although subject-based DLs are beginning to emerge on the Web, and promise us opportunities we never had with traditional libraries or even the Web, DLs in general have not taken off in a "big" way compared to, say the Web. One reason for the Web becoming popular overnight is the introduction of Mosaic, the graphical user interface to the Internet, making it "very easy" for anyone to publish and retrieve information. In contrast, the majority of the current DLs is mainly repositories of information, hence users' experience in DLs is passive and less engaging compared to the Web. Therefore, if the DLs are to be popular with children, they need to be fun, easy-to-use and empowering them — both as readers and authors.

Here, in our prototype, we provide the children opportunity for *creation* of their own stories/poems and uploading them into the temporary DL space for reviews from their teachers and peers, before submitting to the permanent DL (see Figure 1). Only stories and poems approved by the teachers can be submitted to the DL, thus ensuring the quality of the documents.

Figure 1. Children can create their own stories/poems and upload into temporary DL space

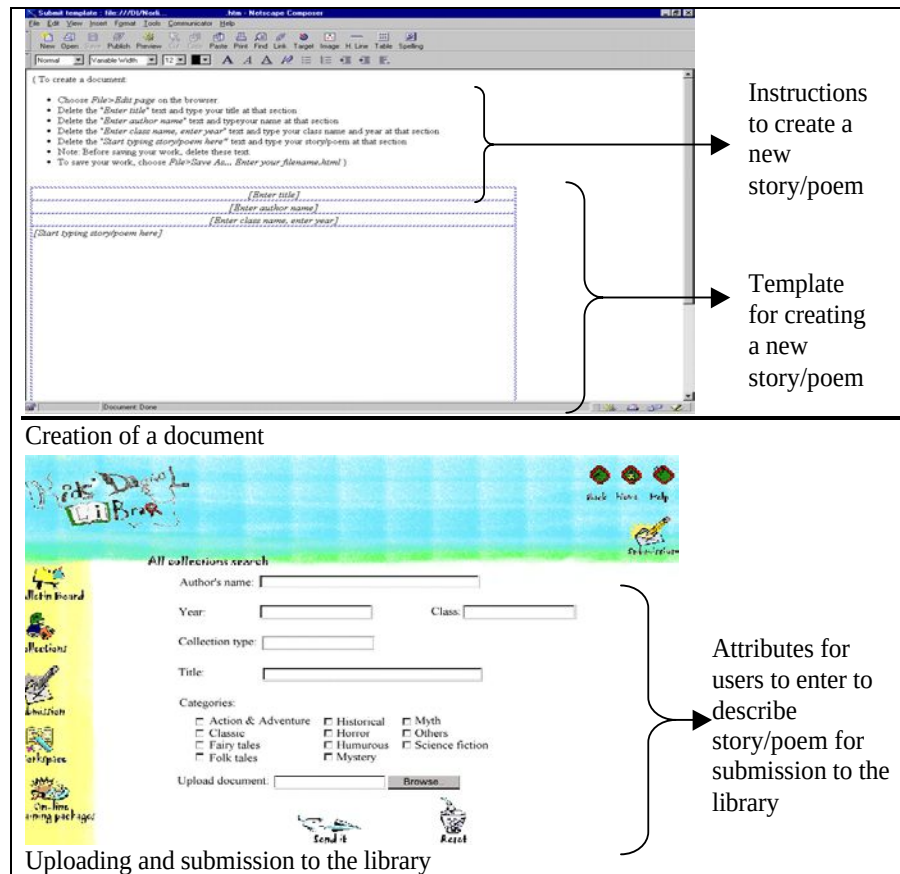
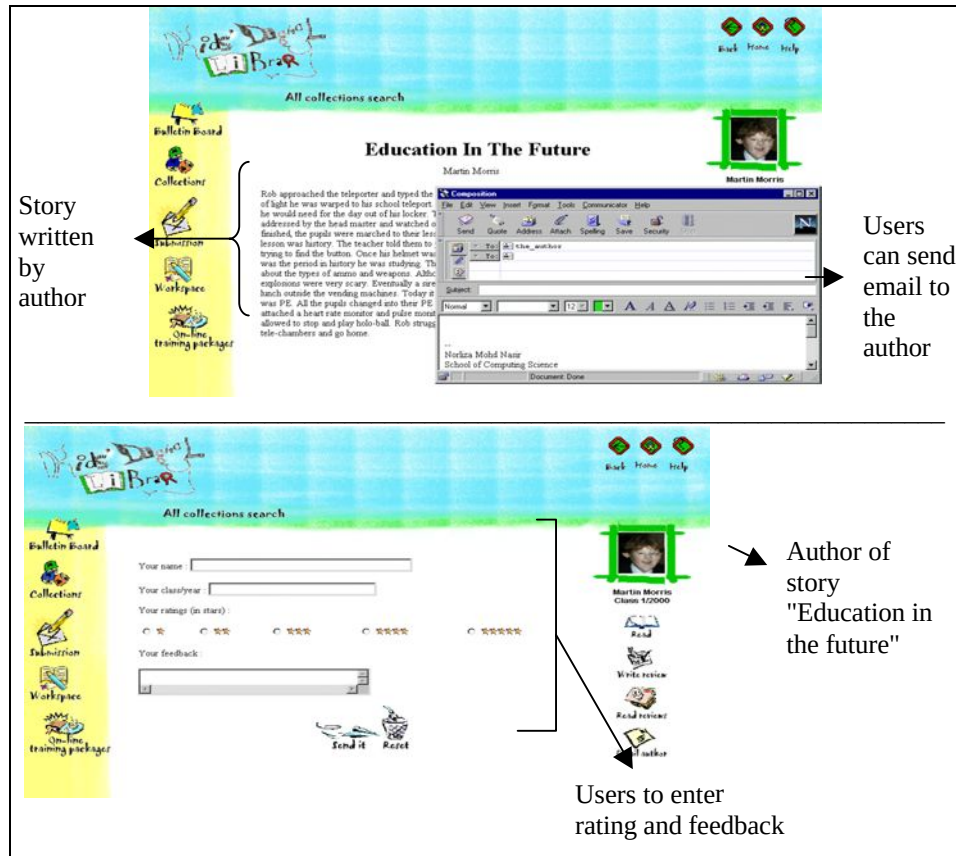


Figure 2 shows one of the stories in the permanent DL. To encourage *collaboration*, children can read stories, give reviews, read other children's reviews on stories and email authors for other comments.

Figure 2. Children can submit review, give feedback and rank story read.



To promote active *engagement* of the children in using the DL, we apply Ellington *et al.*'s four basic factors to match the natural learning processes of humans by [3]: (F1) making materials stimulating and interesting enough to make learners *want to learn*; (F2) incorporating sufficient activities to help learners experience *learning by doing*; (F3) providing sufficient channels of *feedback to learners*; and (F4) enabling learners to *digest and relate* what they have learned to the real world.

In our DL environment, children can query and browse stories and poems written by other children (applying F2). The distinctive feature is that it also allows children to create and submit their stories and poems to a temporary workspace permitting others such as their teachers and peers to read and give feedback by sending their comments via email to the children authors (applying F2, F3, F4).

The DL environment also provides: (1) a display of the top ten stories/poems (applying F1); (2) information about the authors (applying F1); and (3) message board

to post and discuss ideas (applying F3, F4). Resources to help the children authors include links to: (a) other related DLs; (b) helpful on-line dictionary, thesaurus and spell checker; and (c) relevant computer-based training packages to teach them skills on, for example, how to write better (applying F2).

3.2 Provide Workspace and Equal Opportunity for Flexibility and Manipulation of Search Results

We introduce to the DL facilities a "desk" for interaction on the basis of Equal Opportunity, a heuristic introduced for human-computer interaction [11]. Here, the user can exploit the prior output of the computer as input to a further stage in interaction, with or without modification. It may be useful to compare this to the opportunity of activity in a traditional library. Readers are able to get a list of possible items of interest, and retrieve them for further inspection. Those that prove of cursory interest can be set aside or returned to their usual place quickly, and those of greater use can be gathered together for deeper investigation.

Search items can be ordered and prioritised according to either title or author. We are in the process of extending our DL prototype the possibility of allowing search items to be ordered and prioritised according to either the year of publication or by the collection or categories within the library space (see Figure 3).

Figure 3. Search items can presently be ordered/prioritised according to either title or author



4 Participatory Evaluation—Engaging Children as Testers

At the end of the second session, we were certain we wanted the DL to be fun as suggested by the children. This is in agreement with approach encouraged by [4] that designers when designing technologies for children can also make things that ask laughter, excitement and creativity. We were, however, uncertain as to what the children meant by making the DL useful and fun. The children's concept of fun was to include games such as having books falling from shelves and only the relevant books would be opened and "caught" to be read. Other suggestions seem somewhat violent which include "beating up the librarians" or "drawing graffiti in the library" that have no relation to the functionality of the DL (at least to adults!).

Hence, we decided to have a third session with the children to give them something more concrete to comment on their likes and dislikes of DLs, as some children may have a more difficult time verbalising their thoughts, especially when it concerns abstract concepts and actions [12]. The third session was held in February 2000. The children were given three different, paper-based versions of interface designs of the DL but each with similar basic design activities and screen layouts suggested by them (see Section 3). The three different interface designs are:

- *Interface A* using the New Zealand School Journal¹ interface with only the images and icons changed to something more brightly coloured to appeal to a younger user group;
- *Interface B* uses a library metaphor with children moving among book shelves looking for relevant books, articles, etc.; and
- *Interface C* uses a game-like metaphor of an adventure story engaging children in a treasure hunt for books.

Due to constraint of space in this paper, only the design features implemented in Interface A are shown in Figures 1-3. Please see <http://www.cs.mdx.ac.uk/dl/DLinterfaces> for designs of Interfaces B and C.

The aim of the third session was to carry out *participatory evaluation, engaging children as testers* to:

- conduct an initial evaluation of the three different interface designs to investigate useful design features for DLs for children; and
- rate in order of importance design features crucial to the success of children's DLs.

4.1 Initial Evaluation of Three Interface Designs

Protocol of third session (Activity One)

The session began with Activity One showing the children a web-based, mock-up of the three interface designs of the children's DL. The children were then divided into three groups of six to eight with an adult to ensure that the children understood what needed to be done as well as to facilitate discussions if necessary. The groups were given a paper-based version of the three interface designs, allowing the children to examine more closely the different interface designs of the DL. The children were each given a worksheet to write down three things they liked about the interfaces and three things they wished the interfaces would provide.

Results and Analyses

The aim of the third session was to get a quick impression of the children's responses to the "look-and-feel" of the three different interface designs. Table 2 shows things children generally liked about the three different interface designs.

With regard to screen display, the children loved the use of bright colours, good graphics and audio. The simple design with easy access to help in the form of a helpdesk/librarian was well-considered. To them, good navigation involves ease of use to get information and help. Links to other relevant sources and writing aids are

¹ URL for New Zealand School Journal is <http://www.nzdl.org/>

welcomed. Interface B with a library metaphor seemed to appeal to the children because it was more "realistic, easy to use since most people have been to traditional libraries".

Although the children liked the use of colours as reflected in Table 2, they were not keen on having too much of one colour, for example, Interface A having too much blue or Interface B having too much brown. Simple layout is important with appropriately sized text and labelled buttons. Navigation should be kept simple with not too many clicks to find the relevant information.

When we started with this project, we were uncertain as to the likes and dislikes of this age group (11-14 years old) as most research does not work with children of this age group. What was surprising in their comments was that they wanted the interfaces to be "less babyish" and "more adult". It would seem that they preferred interfaces that were fun and interesting to use as well as functional.

Table 2. Things children liked about the three interface designs.

Design Categories	Interface A (conventional DL interface)	Interface B (library metaphor)	Interface C (game metaphor)
<i>screen display</i>	- bright colours - good graphics/icons - simple design - audio - easy access to help - easy to understand	- bright, bold colours - simple terms and presentation - books on shelves - signs on the wall - notes on the noticeboard - helpdesk/librarian	- good colours - good graphics - good idea with message popping up and animation
<i>navigation</i>	- easy to use - easy to get info - links to other sources	- quick access to other items - feedback system	- easy to use - links to other writing aids
<i>overall comments</i>	- creative writing	- colourful and interesting - very realistic, very easy to use because it's like a normal library which most people have been to	- treasure map makes things easier for us to use

4.2 Ranking Design Features for Children's Digital Libraries

Protocol of third session (Activity Two)

Activity Two was to get the children to rate a list of fifteen design features on a scale of 1-4, with number 1 being "not important" and number 4 being "very important". The aim of this activity was to rank in order of importance the design features considered as crucial in the DL by the children in Activity One.

Results and Analyses

Table 3 shows the percentage of children who rated the features as either "important" or "very important". Top of the list is the "ease of use to find information" (96%)

followed closely with the need to make the DLs "fun/interesting/attractive/appealing to use" (91%). The provision of some kind of "librarian-like" help features (for example, guided tours) did not seem to appeal to them, scoring a low 17%.

Although the division between usability and usefulness is blurred as evidenced by differing views held by researchers [e.g. 8, 10, *etc.*], in this paper, we define "usability" of DLs to be measured by established usability dimensions covering usability defects such as screen design, terminology and system information, system capabilities and user control, navigation and task completion [9]. In contrast, "usefulness" of DLs is measured in reference to system specifications and not on user performance testing [9]. From children's feedback (see Table 3), there seems to be an equal emphasis for the children's DL to be usable (53%) as well as useful (47%).

Table 3. List of design features for children's DLs ranked by children in percentages

Criteria	Category	Percentage
1. Easy to use to find information	Usable	96%
2. Fun to use	Useful	91%
3. Interesting to use	Useful	91%
4. Attractive/appealing to use	Useful	91%
5. Easy to retrieve/download materials	Usable	87%
6. Easy to use search facilities	Usable	78%
7. Easy to learn how to use	Usable	74%
8. Simple language, easy to understand	Usable	70%
9. Help feature provided	Useful	65%
10. Relevant to what you are doing	Useful	65%
11. Flexible to use	Usable	61%
12. Links to relevant materials	Useful	61%
13. Feedback given on what you have asked the computer to do	Useful	52%
14. Feedback/hints on what to do next	Usable	48%
15. "Librarian-like" help features (e.g. guided tours)	Usable	17%

5 Design Lessons, Conclusions and On-going Work

To ensure we design a children's DL with and for children, we carried out a study to investigate useful design features children's DLs should have by engaging children as *design partners*. While children can be extremely honest in their feedback and comments, much of what they say needs to be interpreted within the context of concrete experiences [2]. Therefore, in our study, we also involved the children as *testers* providing them with three different interface designs based on what they suggested in the first two sessions. The children gave us useful suggestions congruent to other research findings with younger children [e.g. 2, 4, 5, *etc.*] that children prefer systems that are fun to use, interesting and appealing. As a result, we implemented in our prototypes features for children to create stories/poems to the DL, write reviews, give feedback and to collaborate with other children. In contrast to current DLs, our DL provides a dynamic environment to encourage active engagement of children in using the DL. Unlike the younger children, this age group, however, seemed to want systems that are more functional and "adult in appearance".

In our study, we also wanted to investigate which design features in DLs were considered more important to the children. Contrary to popular belief and other

research findings [e.g. 2, 4, *etc.*], although our children prefer systems to be fun, interesting and appealing, usability of DLs is considered even more important. When we designed the three different interfaces, we thought the children would be more satisfied with the game-like interface. However, results from our study showed otherwise. It will be interesting to repeat our work with other age groups and control for other factors such as web skills and gender.

This is on-going work for us. The initial work we have done suggests useful work with children and DLs. Next, we will be carrying out more in-depth, one-to-one sessions with the twenty-three children to evaluate the usability and usefulness of the three prototypes. The design features discussed in this paper need to be further refined and tested with different types and numbers of children before they can emerge as guidelines for the design and evaluation of digital libraries for children.

Acknowledgements

We would like to thank the children at St. Albans School (U.K.) for their enthusiasm and ideas, and the University of Waikato for the use of the Greenstone software (see <http://www.nzdl.org>). This work was carried out as part of a UK's EPSRC grant (GR/M72098).

References

1. Druin, A.: DLs for children: Computational tools to support children as researchers. <http://www.cs.umd.edu/hcil/kiddiglib/>. (2000).
2. Druin, A., Bederson, B., Boltman, A., Miura, A., Knotts-Callahan, D. and Platt, M.: Children as technology design partners, in Druin, A. (ed.), *The Design of Children's Technology*. Morgan Kaufman. pp. 51-72. (1999).
3. Ellington, H., Percival, F. and Race, P.: *Handbook of Educational Technology*. Kogan Page. (1995).
4. Hanna, L., Ridsen, K., Czerwinski, M. and Alexander, K.J.: The role of usability research in designing children's computer products, in Druin, A. (ed.), *The Design of Children's Technology*. Morgan Kaufmann. pp. 3-26. (1999)
5. Heller, S.: The meaning of children in culture becomes a focal point for scholars. *The Chronicle of Higher Education*, pp. A14-A16. (1998).
6. Kafai, Y.: Children as designers, testers and evaluators of educational software. in Druin, A. (ed.), *The Design of Children's Technology*. Morgan Kaufman. pp. 123-145. (1999).
7. Koenemann, J., Carroll, J., Shaffer, C., Rosson, M. and Abrams, M.: Designing collaborative applications for classroom use: The LiNC Project, in Druin, A. (ed.), *The Design of Children's Technology*. Morgan Kaufman. pp. 99-122. (1999).
8. Landauer, T.K.: *The trouble with computers: Usefulness, usability and productivity*. MIT Press. (1995).
9. Lingaard, G.: *Usability testing and system evaluation: A guide for designing useful computer systems*. Chapman & Hall. (1995).
10. Preece, J., Benyon, D., Davies, G., Keller, L. and Rogers, Y.: *A guide to usability: Human factors in computing*. Addison-Wesley. (1993).
11. Theng, Y.L., Duncker, E., Mohd-Nasir, N., Buchanan, G. & Thimbleby, H.: Design guidelines and user-centred digital libraries, *Third European Conference ECDL'99*, Abiteboul, S. and Vercoustre, A.(Eds.), pp. 167 - 183, Springer. (1999).
12. Underwood, J. and Underwood, G.: *Computers and Learning: Helping children acquire thinking skills*. Blackwell. (1995).

Evaluating a User-Model Based Personalisation Architecture for Digital News Services

Alberto Díaz Esteban¹, Pablo Gervás Gómez-Navarro¹, and Antonio García Jiménez²

¹ Departamento de Inteligencia Artificial, Escuela Superior de Informática, Universidad Europea-CEES

Villaviciosa de Odón, Spain
{alberto, pg2}@dinar.esi.uem.es

² Departamento de Periodismo Especializado, Facultad de Ciencias de la Información, Universidad Europea-CEES

Villaviciosa de Odón, Spain
antonio.garcia@fcp.cin.uem.es

Abstract. An architecture that provides personalised filtering and dissemination of news items is presented. It is based on user profiles and it provides mechanisms that allow the user to control and tailor to his own needs the interaction between three different sources of relevance judgements: the existing newspaper categorisation by sections, basic information retrieval on user selected keywords, and an additional operation of automatic categorisation against an alternative hierarchy of categories. These three tiers cover some of the most promising access methods for digital libraries. The proposed architecture has been implemented and evaluation results are presented, covering user response, system efficiency, and user preferences regarding the set of methods made available to them.

1 Introduction

Personalised information services are becoming popular in the Web. Many major digital newspapers around the world now offer to send users by electronic mail a customised selection of news items. In all the instances surveyed by our research team, the selection of information is carried out by simple methods: users choose which sections of a newspaper they are interested in or certain key words that they would like to appear in their selected items. We have performed an evaluation on 15 news services, mainly Spanish, to extract conclusions about the current state of the art. Based on this study we have constructed a system that incorporates these basic methods, together with existing techniques in the fields of natural language processing and user modelling, into the process of selecting the particular information items that are relevant to a given user. The system allows readers of the newspaper to receive a periodic e-mail message containing the news items that the system finds particularly relevant to the interests of the user, previously defined during registration. The resulting

architecture provides an additional control layer that allows users to specify how important each of the different methods of selection is to his particular interests. In this way, the system constitutes an integrated and customisable option of combining document relevance information from sources of three different kinds:

- A prior categorisation by the system owner (documents, i.e. news are sorted into sections by the editor)
- Keyword search information over the content of the documents
- Automatic categorisation results against an alternative (less domain specific) set of categories

As such, it presents a flexible, multidimensional and browsable user model, and a set of well founded techniques for user models and information items matching. The user model captures the user information interests by means of complementary aspects like newspaper sections, categories extracted from a general purpose search engine, and keywords extracted from read news articles. The integration of text retrieval [7] and categorisation [4], [5], [11] based on the Vector Space Model provide a solid theoretic framework for the implementation of the service.

2 Analysis of Other Systems

We looked for systems that provide news service on the Internet to be evaluated [9]. The following systems were selected: Crayon, Diario de Navarra, Economyweb, El Diario Vasco, El Heraldo de Aragón, El Mundo Digital, El País Digital, Expansión, Hispanidad, La Razón, La Vanguardia Digital, Politicsweb, Telépolis, The Economist, The New York Times.

Each system was monitored over a period of time by a team of four research students, performing a quantitative and qualitative analysis. Following [6], a quantitative analysis was carried out in which a form of 103 questions was filled out by each evaluator for each system. The questions were grouped around four specific topics: interface evaluation, categorisation evaluation, summary evaluation and actual relevance of received documents. For each of these parameters a numerical value was worked out from the evaluators responses. In each case a total percentage value was obtained for each system. Qualitative analysis was based on the description and valuation of each system, and personal interviews with the evaluators describing their overall impression of the different systems.

The following conclusions regarding the systems themselves arise from the evaluation: a) evaluators consider it important for systems to provide categories and keywords as means of selection (a second level of categories is considered a plus); b) systems that allow some personalisation rate much higher than others; c) personalisation should allow a choice of when to receive information and in what format; d) the effectiveness of some systems is low; e) access to the final documents must be provided; f) including a summary is a key positive feature; g) the presence of unrequested or irrelevant information (noise) can severely spoil otherwise good ratings; h) users dislike systems that provide too large a number of items; i) subscribing and unsubscribing must be easy tasks; j) ease of access and use is important.

All these results were taken into account when designing the personalisation architecture described below. The questionnaires and general evaluation method employed were later applied to the resulting prototype.

3 Modelling the Users

After the study of some commercial systems we propose a browsable user model designed to represent the user information interests in a wide variety of ways. This describes the user information needs, that is, what a user is really looking for.

Of the ingredients of a user model proposed in [1], for this particular application information type, document structure, means of delivery and source of the information need not be modelled, since these aspects are invariable in a typical digital news service.

The user model stores three main kinds of information:

- The personal information of the user, which includes name, login, password and email address.
- The format information for the messages, which includes the weekdays the user wants to get a message, an upper bound on the number of news items per message, and an “on holidays” binary value (which allows to put the system on hold for specific periods of time).
- The information about the user interests, which covers newspaper sections, general categories, and keywords.

The format information is very useful for the users. Both imposing an upper bound on the number of items per message and the possibility of putting the system on hold avoid message overloading. Establishing a lower bound has been considered counter-productive, since it may lead to the inclusion of noise in messages whenever the lower bound cannot be met with relevant information.

As described in section 2, news services usually support personalisation by means of sections and/or keywords selection. Our model supports both features:

- The users can select their preferred sections. The sections of a newspaper are a primitive set of content-based categories. This set of 9 different categories comes from the traditional organisation of a newspaper. Users can assign a weight to each section, in order to prioritise the news items coming from it. Examples of sections in the newspaper are “International” or “Culture”.
- Users can also provide a set of chosen keywords by typing them in. Each keyword provided has to be given a weight that represents its relative importance for the users interests.

The system allows users to edit their selections: to modify the weights for sections or keywords, or to add or remove keywords.

As an additional selection feature, an alternative classification of the news items, obtained by means of automatic categorisation of the documents against a different set of categories, is provided. Internet users are already familiar with the categories systems employed in search engines directories. The category system of Yahoo! Spain was chosen as an alternative way of representing interests. Since it is designed for a

wider purpose and a wider domain than newspaper section headings, it constitutes a good second opinion. The first level of 14 categories from Yahoo! Spain is presented as a choice. Users can assign a weight to each category, to represent their interest in it.

Too many methods of selection available simultaneously can lead to confusion. Unless additional control features are provided users get at most a blurred picture of the operation of the system. For this reason, our personalisation architecture allows an extra level of user specification. A general control mechanism has been included to make the results more predictable for the user. Each of the three features (keywords, sections and additional categories) has a weight that represents its importance for the user interests. For example, if the weight of sections is low and the weight of additional categories is high, relevance values concerning additional categories will be considered more important for selecting news items.

In this way, each of the three dimensions considered in the user profiles can be defined and controlled by the user, providing a fine-tuning mechanism to obtain a flexible characterisation of his interests.

4 Information Filtering

The user model is applied daily by the system to the news items of the day, using Text Classification techniques (Text Categorisation and Information Retrieval [2], [4], [3]). A ranking of the news items is obtained according to their relevance for the given user. The top of the ranking is selected for the user in accordance to the upper bound on number of items per message specified in this profile.

4.1 Representing Information

The news item are downloaded daily from the newspaper website in the form of html documents. The title, section, URL and text for each document are extracted and stored. The representation of the news is obtained applying Vector Space Model (VSM) [7] to the text.

The VSM was originally developed for Information Retrieval (IR), but it provides support for many text classification tasks. The VSM for IR is applied by representing natural language expressions as term weight vectors. Each weight measures the importance of a term in a natural language expression, which can be a document or a query. Semantic closeness between documents and queries is computed by the cosine of the angle between document and query vectors. The terms selected for the representation are those that do not appear in a stop list for Spanish. We use the formulae based on term frequencies to compute their weights.

A representation for each category can be obtained by applying text categorisation techniques [4], [5] and using a set of training documents¹. In our case, the set of training documents used were the web pages indexed by Yahoo! within these categories.

¹ A set of documents labeled manually with the suitable categories.

Thus, each category can be represented by a term weight vector that is obtained from the name of the category, the name of its subcategories, if there are any, and the names and short descriptions of the web pages associated to the category.

The keywords also are represented with VSM, using the weight assigned for each word in the model.

4.2 Integrating Text Classification Tasks

We applied Text Categorisation using category-pivoted categorisation [4], [5], [8] with the categories against the news to obtain a ranking of the different news ordered by relevance for each category.

We applied also Information Retrieval [3] with all the keywords against the news to obtain a list of relevant documents for the user.

Also all the news are processed to check if they belong to one of the sections selected in the user model.

When all the documents have been sorted according to the different sources of relevance, the resulting orderings are integrated by using the level of interest that the user assigned to each of the different reference systems. This implies that users looking for the same information but having chosen different methods to specify their interest may get different results. For the relevance values provided to the user to be easy to interpret, they are normalised over the number of selection methods involved in obtaining them. In this way, the system can quote a final relevance value in the range 0-100% to every user regardless of the number of selection methods that he chose.

4.3 Information Dissemination

A message is generated for each user with the selected documents respecting the preferences stated in their profile. The message is sent by email early in the morning if the current day has been selected in the profile. The exact time is chosen to guarantee that the news of the day have been just placed at the disposal of the public in the corresponding web page. The user receives a message that matches the newspaper's design.

The automatic summary is generated by extracting from the HTML document the sub-heading, which tends to be a short summary in itself, provided by the editor. If there is no sub-heading, the first paragraph of the document is extracted. This heuristic gives good results due to the typical structure of news features, where the key ideas are presented at the beginning of the document.

A message is composed of: a) Title of the message with the current date and the name of the user; b) A link to the user model to permit change it if the user wants ; c) Various links to the newspaper (homepage, sections..); d) Brief description of the interests of the user (as featured in his profile); e) The selected documents, presented ordered by relevance and respecting the upper bound selected by the user (for each one: title, hyperlink to the original document, name of the section that it belongs to, final relevance value obtained, short automatic summary of the document, extra link explicitly stating it allows access to the full document).

5 Evaluation

We describe and discuss the three kinds of evaluation that were carried out: an evaluation carried out by a set of different users, a system evaluation that considers the performance of the system in measurable parameters, and an evaluation of the user model provided and how the evaluators have fared in dealing with it.

A controlled evaluation environment was established to allow analysis of the results with respect to the different kinds of user involved. Evaluation was carried out by 44 users in four categories: A) Collaborators; B) Researchers; C) University lecturers (both on Computer Science and Journalism); D) External users with no professional relationship with the fields involved.

The system was evaluated following the working pattern applied to other existing systems (see section 2). For the relevance of the received documents the users had to check the performance of the system against the actual set of documents available on the newspaper website on three particular days. Additionally, on those particular days, logs of system operation (available documents, user profiles at the time, and system selections for each user) were kept to allow objective results to be obtained. With this data we worked out two sets of recall and precision figures: one based on user criteria as put down in the forms, and one based on subsequent close analysis of system logs.

5.1 User-Centred Evaluation

During the first stage, the evaluation was centred on user response and the vision that users develop of the system. The aim was to harvest explicit evaluations provided by the users about system response-time, ease of use, system efficiency, and conceptual and physical presentation. This information was compiled on the basis of a closed questionnaire with specific questions on the relevant main topics. For each of these parameters a numerical value was worked out from the users responses.

In general, users found the system suitable although had some differences between different groups of users. This were the results for the interface evaluation: System Access: (high); General Interface, User Adaptation, and Integration into User Environment: (medium-high); Management of Contents, Query and Retrieval Schemes and User Help (medium). With respect to newspaper sections the following results were obtained: Expressive Faithfulness, Objectivity and Relevance (high). With respect to categories the following results were obtained: Expressive Faithfulness and Objectivity (medium-high); Relevance (medium). With respect to summaries the following results were obtained: Summary Content (high); Summary Structure (medium-high).

Recall and Precision rates have been estimated based on user impressions (see table 1), under the assumption that aim is not to obtain conclusive results but to draw roughly significant conclusions.

Group	Precision	Recall
A & B	0.9	0.8
C	0.9	0.6
D	0.9	0.6
Average	0.9	0.7

Table 1. User estimated Recall and Precision (by groups)

Additionally, the qualitative analysis showed that users were satisfied with the system characteristics, personalisation quality, formal quality, and categories system. On the other hand, the users' familiarity with similar systems influenced their understanding of the basic mechanisms. Some users found it could be more visual, but most of them understood it after receiving the first message.

5.2 Evaluation of Observed User Profiles

The analysis of the 44 user models logged with the system yields the following data (see table 2).

	Upper bound	Selection methods	Sections	Categories	Keywords
Average	14,0	1,9	2,6	3,4	2,3
Max	20,0	3,0	9,0	14,0	15,0
Min	5,0	1,0	0,0	0,0	0,0
Selected values	44,0	44,0	30,0	26,0	18,0
Selected average	14	1,9	3,9	5,8	5,7

Table 2. Analysis of user profile development

The average selection of a user has approximately 14 as upper bound of documents per message, 2 methods of selection (in most cases, sections and categories), 3 sections, 3 categories and 2 keywords.

All the users selected the sections method, with or without other method of selection, except one that chose to use only categories and keywords.

All the users select some method and some upper bound, but not all select all methods. Thirty chose sections, 26 chose categories and only 18 chose keywords. It seems that less intuitive methods are less favoured. The users that chose the sections method choose an average of 4 sections. Those that chose categories, marked 6, and those that chose keywords, marked 6. When the user opts for a method, he tends to select more than one possibility.

Some user select a method (section information for instance) but do not select any particular criteria for it (mark no specific sections). This results in an empty user model. This issue appeared for 14 users, all of which chose only sections. This has been identified as a problem that needs further work.

Regarding differences in profile development between user groups, it has been observed that groups A and B (which had taken part in the development of the project) tended to restrict their selections more than groups C and D: the number of selected sections and categories on average rose steadily from A to D.

5.3 System Evaluation

We computed the values of recall and precision and others features for all the users on the last of the three specific days that the user had to review exhaustively. This allows a comparison between user evaluation and system evaluation to check the exhaustiveness of the user judgements and check the true performance of the system.

We have obtained the following results for each user: recall, precision, number of news selected by the information filtering system, that is, news with relevance greater than zero, number of truly relevant news. The day of this evaluation had 108 news.

Table 3 shows the average results and the maximum and minimum values for each feature.

	Upper bound	Relevant news	Recall	Not relevant news received	Precision	News selected
average	14	69,2	0,2	0	1,0	96
max	20	88	1	1	1,0	107
min	5	11	0,1	0	0,9	11

Table 3. Recall and Precision figures from system logs

5.4 Results Discussion

Studying the results we can see that our system refines the information of the sections with the categories and keywords. The average precision is close to one because the fact that a document belongs to a section is enough for the document to have a high relevance value. Moreover, the relevance for belonging to a section is always greater than the relevance for belonging to a category or containing a keyword. Since most users have selected at least two sections, a section holds an average of eleven documents, and the average upper bound of documents per message is 14, most users get messages where all selected documents are relevant.

If a user marked sections as selection method in his profile (most do, in fact 50% of the users rely on sections altogether to select), the selected documents that appear first belong to these sections. They are shown ordered according to relevance computed with respect to categories and keywords. They are followed by documents that do not belong to these sections but are relevant in terms of categories and keywords. These show much lower relevance values nonetheless.

However, a user that does not use sections obtains documents sorted by the information relative to categories and keywords, and so obtains relevant documents from different sections. Only one user operated in this way, and he obtained similar value of

precision and a low value of recall. This is because he had selected 7 categories and 14 keywords and the number of relevant documents under such wide criteria is above average.

If the relevance value computed using the categorisation method were greater than it currently is, documents relevant according to this source might find their way to the top of the ranking. This is at present unlikely because the categorisation system yields always very low relevance values, but we hope to improve it by developing a richer representation for categories.

If we compare the results of the user evaluation and the system evaluation we can see that the precision obtained is very similar but the recall is lower in the system evaluation. The reason is that a user considers a document as relevant if it refers to something that is interesting for him, whether or not it belongs to a category or contains a word. However, the low recall value is a consequence of the upper bound imposed by the user: with a user model with a few sections and few categories the number of relevant documents is too high to be captured in a maximum recall fixed for the user by the upper bound.

6 Conclusions

We have constructed a system that implements a personalisation architecture based on a complete user model that captures different interests of a user. The system operates in the domain of digital news services. It improves on existing personalisation methods by relying on the user model and text classification techniques to filter the particular information items that are relevant to a given user.

Based on an evaluation of 15 digital news services the following features have been found desirable: use of more information than sections or key-words to describe the user's interests, effectiveness, personalisation, ease of use of the system and absence of noise. We performed an evaluation of our proposal in three ways: evaluation by users, evaluation of observed user profiles and system evaluation.

The results show the improvements of our system with respect to the other systems analysed: a better personalisation through a more complete user model that integrates text classification techniques. They have also shown that more careful guidance of the user is needed (help documentation, contact persons). It has been noted that users have difficulties in filling out the model in a correct way, generating bad profiles in some cases. For optimal use, the system should provide specific instructions about which method is better for each kind of search. In particular, more refined ways of integrating the different methods must be explored; and evaluation processes that allow some means of measuring the effect of integration should be developed.

The technology employed is not domain specific and relies solely on general techniques. Therefore it can very easily be ported to other domains where there is an important volume of new textual information to be processed periodically, like digital libraries where we can use the same method on articles of an author, filtering the documents that belong to a category and with some keyword.

As lines of future work, we are considering: the possibility of creating personal categories to allow the users to define extra categories suited to their own information needs, the use of relevance feedback [11] in our system, a second level of categories to get a more specific model, the use of a stemmer for Spanish and Spanish Wordnet to take into account different lexical relations between words, and extending our system to a multilingual framework by taking advantage of Eurowordnet [10] and other multilingual resources.

References

1. Amato, G., Straccia, U. (1999). User Profile Modelling and Applications to Digital Libraries. Third European Conference, ECDL'99, Paris, France, September 1999.
2. Díaz, A., Buenaga, M., Ureña, L. A., García, M. (1998) Integrating Linguistic Resources in an Uniform Way for Text Classification Tasks. First International Conference on Language Resources & Evaluation, Granada (Spain), 1998.
3. Frakes, W. & Baeza, R. (1992), Information retrieval: data structures and algorithms, Prentice Hall, London. 1992.
4. Gómez Hidalgo, J. M. and Buenaga, M. (1997). Integrating a Lexical Database and a Training Collection for Text Categorisation. ACL/EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP, Madrid (Spain), 1997.
5. Lewis, D.D., Schapire, R.E., Callan, J.P. & Papka, R. (1996). Training algorithms for linear text classifiers. In Proceedings of the ACM SIGIR, 1996.
6. Pastor, J. A., Asensi, V. (1999). Un modelo para la Evaluación de Interfaces en Sistemas de Recuperación de Información, IV Congreso Isko-España Eoconsid'99, Granada (Spain), 1999.
7. Salton, G. (1989). Automatic Text Processing: the transformation+, analysis and retrieval of information by computer. Addison Wesley. 1989
8. Sebastiani, F. (1999). A Tutorial on Automated Text Categorisation. Proceedings of the First Argentinean Symposium on Artificial Intelligence (ASAI-99)
9. Theng, Y.L., Duncker, E., Mohd-Nasir, N., Buchanan, G., Thimbleby, H. (1999). Design Guidelines and User-Centred Digital Libraries. Third European Conference, ECDL'99, Paris, France, September 1999.
10. Vossen, P. (1996). Eurowordnet: building multilingual wordnet database with semantic relations between words. technical and financial annex. Technical report, EC-funded project LE # 4003.
11. Yang, Y. (1999), "An Evaluation of Statistical Approaches to Text Categorisation", Information Retrieval Journal

Aging Links

Claudia Niederée, Ulrike Steffens,
Joachim W. Schmidt, and Florian Matthes

Software Systems Institute,
Technical University Hamburg-Harburg,
D-21073 Hamburg, Germany
{c.niederée,ul.steffens,j.w.schmidt,f.matthes}@tu-harburg.de
<http://www.sts.tu-harburg.de>

Abstract. Rooted in the principle of hypertext, linked information is almost ubiquitous due to the WWW and related services. Links between information objects are established for various reasons, aspects of which are encoded by a link type or expressed through link context, e.g., by the surrounding content. Such reasons may lose their validity through content evolution in the link target. Fine-grained solutions are required that enable the user to gain evolution awareness without being distracted from his main task.

In this paper we present *aging links* as a non-intrusive mechanism for improving awareness in cooperative work with linked information networks. A link may age, affected by the evolution of the link target, leading to a gradual loss of its validity. The aging process is driven by evolution-indicating events and may be flexibly controlled by link type specific aging strategies.

A customizable service, *EvEnAge*, based on standard technologies, prototypically implements the concept of aging links for XML documents.

1 Introduction

In selecting, collecting, and structuring relevant information objects from the universal information space, cooperative work with linked information is an important activity in modern information environments such as digital libraries, intranets, and portals. Linked information networks are subject to permanent content evolution where the user community is strongly involved in such changes.

The rich semantics of linking is based on the manifold reasons for link establishment and the high degree of autonomy on both sides, the link target with its autonomous content evolution and the link source with its individual content use. If the reasons for linking are made explicit in the context of the link, e.g. by the surrounding text, the evolution in the link target may corrupt the correctness of the information object containing the link source. Thus, link validity depends greatly upon the rich and varying link semantics. Fine-grained solutions are required to enable users to gain awareness [5] of the evolution-related aspects of link validity without being hindered in using linked information networks.

For this purpose, we propose the concept of aging links. Increasing ages are assigned to links reflecting the effects of evolution in the information object targeted by the link. The aging process is

- driven by evolution-indicating events like link target version creation;
- controlled by link type specific aging strategies influencing the speed of aging;
- consolidated by the introduction of common age levels that create a uniform notion of age and facilitate a comparison of link ages.

In this paper we take the position that link aging contributes to evolution awareness by visualizing the evolution of linked information objects and by triggering cooperation- and evolution-related actions on the basis of link ages.

The paper is structured as follows: Section 2 motivates our approach by discussing the benefits of link aging and its exploitation in different cooperative scenarios. Section 3 introduces the concept of aging links. It describes the aging process based on evolution-indicating events that drive this process. *EvEnAge*, a web-based service implementing aging links, is presented in section 4. The paper concludes with a discussion of related work and future research directions.

2 Evolution in Cooperative Link Scenarios

Being part of an information network the modification of an information object such as a text document, an image, etc., may not be viewed in isolation but also its impact on the objects by which it is referenced. We propose the concept of *aging links* to cope with this situation. Driven by the evolution of the target object, a link ages, thereby reflecting a (possible) decrease in the validity of the link. Link aging contributes to evolution awareness in a wide variety of cooperation situations. In the following sections, a web-based competence center is used to illustrate and discuss the benefits of link aging.

Link Aging in Loose Cooperation Scenarios. A web-based competence center is established to manage and distribute knowledge among the members of an organization or a community. Typically, it contains numerous annotated links to related information objects from different providers. This may for example be a link to a standard specification (e.g. of XML) augmented with a comment on its relevance. The validity of such annotated links is often highly evolution dependent. The annotation on an external JDBC user guide may state that it does not cover JDBC 2. A new version of the guide may introduce this feature and invalidate the annotation. This decrease in validity can be reflected by aging the respective link.

Due to the distributed situation in this loose cooperation scenario, link aging requires evolution protocols between the competence center that acts as an information consumer and the external information providers.

Link Aging in Close Cooperation Scenarios. In the process of cooperatively constructing the information artifact “competence center”, information objects are collected, composed, classified, and linked to internal as well as external resources.

Parallel work is often unavoidable when teams work creatively with information. Two groups may, for example, work on an introduction to Java programming in parallel, both tackling the problem from different backgrounds. The results of such parallel work are later merged or kept as alternatives for a

different readership. A link stating that some information object is an alternative to another one is rather stable with respect to changes in the objects. An adequately customized and type-dependent aging strategy reflects this fact.

Working with two copies of the same information object a link may, on the other hand, connect the copy with its master. Link aging informs the owner of the copy about evolution in the master. Here, different degrees of autonomy may be implemented via customized aging strategies.

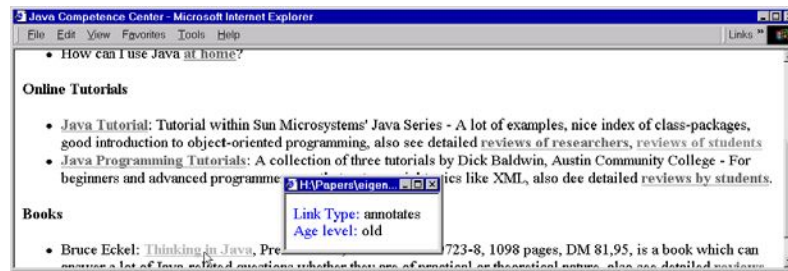


Fig. 1. Rendering Aging Links

Exploiting Link Age. The competence center team may react to aging links by rechecking the referenced information objects and, e.g., adapting the link context to the changed link target. The support of different age levels refines the awareness mechanism: A middle-aged link may still be left alone for a while, whereas a very old link requires action, for example in the form of re-reading the changed information object. A user may choose to be informed about the reaching of a new age level via a changed visualization or an explicit notification. Figure 1 shows the visualization of link aging via fading colors. Further possible exploitations of link age include the retrieval of links that are above a certain age level, e.g. to recheck them, or the automatic deactivation of links exceeding a certain threshold value.

3 The Concept of Aging Links

The metaphor of *aging* is used to reflect the effects of target evolution on a link: A mechanical link ages over time either by its use or non-use, leading to a gradual loss of its functionality, similarly a digital link ages, affected by the evolution of the link target, leading to a gradual loss of its validity.

3.1 The Aging Process

Links in our approach are subject to an *aging process*. This process is driven by rated events indicating evolution in the link target: Each link is associated with a value that reflects its age. This value is set to zero upon creation of the link. Whenever a new version of the link target is created or another evolution-indicating event occurs, the link age changes to reflect the new situation. Update-in-place semantics is assumed here. This may also be achieved by an indirection (see e.g. PURLs [17]).

Aging is considered for typed directed links between information objects. As discussed in [8], typed links provide added value in information networks e.g. by enabling machine supported, link type specific analysis and actions. Exploiting link type specific evolution dependencies empowers a fine-grained control of the aging process.

Age comparability is an important issue for the evaluation of link ages by the user. Unfortunately, links of distinct types differ in their aging characteristics impeding the direct comparison of ages. In order to create a *common* notion of age, so called *age levels* have been introduced together with link type specific mappings from concrete ages to these age levels. This idea was inspired by the use of the terms “old” and “young” in nature: A tree is still called “young” when it is 15 years old whereas a dog aged 15 is already an old dog.

A link traverses a sequence of *age levels* that partition its lifetime into phases of assumed validity. Links in the same age level can be expected to be in roughly the same state of validity, although their targets possibly have been subject to a considerably different degree of evolution. Age levels are associated with type specific threshold values. If the link age reaches a threshold value, the link transits to the associated age level. This transition is exploited to trigger evolution-related actions like changing link visualization or notifying interested cooperation partners.

An example set of age levels is: *up-to-date*, *very young*, *young*, *middle*, *old*, *very old*, and *outdated*. Optionally, a threshold value for the invalidity of a link may be defined: When the link reaches the associated age it is deactivated.

The aging process is controlled by an aging strategy that influences the speed as well as the impact of aging. This strategy determines several parameters of the process, i.e. the supported age levels, their threshold values, and the actions triggered by the reaching of an age level. This enables the adaptation of the aging process to the characteristics of different link types and applications.

Each link is associated with an aging strategy. Three levels of strategies are available: A *general strategy* provides a default setting. Link type specific evolution dependency, as it is discussed in section 3.4, is captured by *link type specific strategies*. The definition of *individual strategies* copes with the semantics of special links. This approach combines user comfort with high flexibility.

3.2 Computing Ages and Age Levels

As part of the aging process an age function computes the new link age. It is the difficult task of this function to judge to which degree the evolution in the target information object has affected the validity of the link taking into account the reason for its establishment. This requirement raises three questions:

What are the metrics for link aging? The aging of a link does not depend directly upon time but is triggered by evolution in the link target. A straightforward approach is aging a link by one unit on each new version of the target object. However, this equivocal treatment fails to differentiate between versions introducing only minor changes and those substantially altering the target object. In our approach, evolution is rated with a value of type real that is used to compute the link age. This also enables aging process variants (see section 3.3).

How to handle and rate evolution? Evolution that effects the validity of the link may be based on changes in the link target as well as on other changes in

the environment of the information network, e.g. the elapsing of a certain amount of time. In our model we rely on events that indicate that some kind of evolution has taken place in the target information object. The rating of the *evolution-indicating events* is based on event classification.

How to take reasons for link establishment into account? The reasons for link establishment are only partly made explicit in the link context, e.g. by surrounding text. Computing the effect of the evolution on the link from this text is forbiddingly complicated because of the high freedom in its composition. Therefore, this aspect is approximated by link type dependent heuristics (see section 3.4). Link types code part of the reason for link establishment.

3.3 Classification of Evolution-Indicating Events

The concept of link aging is generalized by introducing the notion of an *evolution-indicating event*, of which version creation is only a special case. Further types of evolution-indicating events lead to several variants of the aging process.

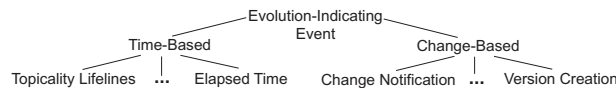


Fig. 2. Classification of Evolution-Indicating Events

According to the classification in figure 2, evolution-indicating events are divided into change- and time-based events. Change-based events may be generated by version creation but may also rely on general change notification. In distributed environments such change notification may be provided by the server managing the target object or by a third-party service like Mind-it [12] that monitors the changes in a wider information space. In contrast, time-based events are triggered when a certain amount of time has elapsed. The impact of an evolution-indicating event on aging may be refined by further classification or parameterizing of the events. Viewing the aging process according to this model allows a smooth integration of time-driven with evolution-driven aging.

A fine-grained rating of events is achieved if versions are classified according to the degree of change they introduce. This can either be performed manually or automatically by difference computation based for example on change detection mechanisms [2] or the domain-based comparison of semistructured documents [10].

The topicality of an information object may vary over its lifetime. By defining "topicality lifelines" for information objects this phenomenon may be exploited as a further source of time-based link aging. For a discussion of this idea see [13].

3.4 Link Type Specific Evolution Dependencies

Link types imply a specific semantic relationship between the source and the target information object of the link. As a consequence, links of some types tend to be more evolution-sensitive than others.

Table 1. Link Type Specific Evolution Dependency

Link Type	Evolution Dependency	Notes
<i>annotates</i>	high	For details see Sect.2;
<i>origin</i>	middle to high	To avoid high divergence between master and copy, evolution dependency is generally defined as middle to high. It may be reduced to increase autonomy.
<i>child</i>	middle to high	The relationship between parent and child concepts is rather strong resulting in a considerable impact of child evolution on this relationship.
<i>alternative</i>	low	For details see Sect.2;
<i>instance</i>	low	A subject of an information object (in our case the basis for classification) is a rather stable property. It is only affected by (large) topic-influencing changes.

A wide variety of link types is possible in information networks [22]. Our choice of five different link types for a closer examination of their evolution dependency was inspired by the experience from two Digital Library projects (WEL [19], Kolibri [18]). Special focus is given to subject-oriented classification hierarchies and to support for cooperation, personalization, and autonomy [16].

Considering the classifiers as information objects of their own, links of type *child* are used to represent classification hierarchies (parent $\succ$ child) and links of type *instance* express the classification of information objects (classifier $\succ$ classified object). A basis for personalization and cooperation support is provided by links of type *annotates* (annotation $\succ$ annotated object) and by links of type *alternative* that avoid premature establishment of consent. A link of type *origin* is used to connect the copy of an information object to its original (copy $\succ$ original).

Table 3.4 summarizes the evolution dependency trends observed for the application domains examined. Deviating evolution dependencies may be found in other domains as well as for individual links requiring an adaptation of aging strategies.

4 *EvEnAge*: An XML-Based Service for Link Aging

The concept of aging links has been implemented by the prototypical aging service *EvEnAge*. This service focuses on web-based applications and enables the aging of links embedded in XML documents. The *EvEnAge* implementation relies on standard technology and is open for integration with existing services and tools.

4.1 The *EvEnAge* Service Architecture

The architecture of *EvEnAge* mainly consists of two components (see Fig. 3): The *Ager*, that computes the ages and age levels for links, and the *Strategy Manager*, that provides access to a repository of aging strategies.

The *Ager* interacts with one or more *evolution controllers* to obtain information about relevant evolution-indicating events. The controller receives these events from components like version managers or third-party change detection services and stores them. The *Ager* also exchanges information with the aging client, where the aging information is stored. The aging links are embedded into XML documents enabling easy integration of link aging into existing web documents and applications.

The *EvEnAge* service is part of a service framework for evolution management in information networks, the evolution engine *EvEn*. *EvEn* contains evolution-related services like notification, version management and strategy-based document merging [11]. This service framework is currently under development at our department.

4.2 Link Ages: Management and Rendering

Representing Aging Links. Avoiding a central storage on the aging server, the aging information is inserted into the client documents. This is achieved by extending XML link elements [4] with additional attributes:

```
<!ATTLIST AgingLink
    xmlns:xlink CDATA      #FIXED "http://www.w3.org/1999/xlink"
    xlink:href CDATA      #REQUIRED      type NMTOKEN #IMPLIED
    age NMTOKEN #REQUIRED      ageLevel NMTOKEN #REQUIRED
    strategy NMTOKEN #IMPLIED      lastUpdate NMTOKEN #REQUIRED>
```

This meta information comprises the link age, the associated age level, the link type and the date of the last age update. In addition, the link element has an optional strategy attribute that identifies the chosen aging strategy. If the attribute is omitted, the aging strategy will be determined by the link type.

Rendering Link Age. Focusing on non-intrusiveness, visualization of link aging by fading link colors was implemented: The color used for the link fades with the reaching of higher age levels (see Fig.1). In case a threshold value

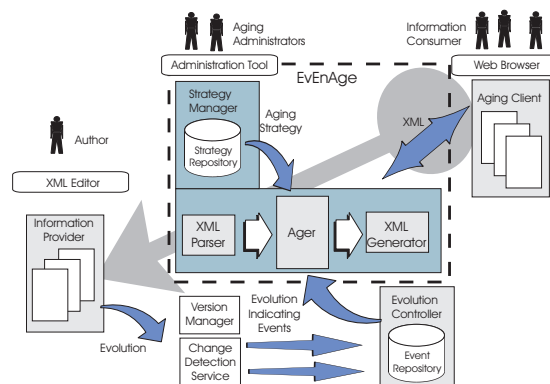


Fig. 3. The *EvEnAge* Service Architecture

for link deactivation is defined, the link color changes to red when the age approaches this value. This color signal means that user interaction is required to avoid losing the link.

Age level visualization is achieved via an XSL stylesheet¹. Hence, age visualization may be easily adapted to user preferences (like colors) and the combination with existing stylesheets is conveniently supported [3]. For the rendering task a web browser environment is sufficient. These properties empower easy integration of the aging concept on the client side.

In future it is planned to generate such stylesheets from the information stored in the aging strategy repository. This enables a dynamic extension and adaptation of the supported age levels and visualization properties.

4.3 EvEnAge Components

The Ager. The *Ager* regularly visits the XML documents of the aging clients and updates the age information. An incremental approach to age computation is implemented by the tool. The client documents are processed by an XML parser [9] that provides the input to the *Ager*. The output of the *Ager* is handed over to an XML generator. The computation of the output, i.e. the new age information, consists of the following six steps:

- Step 1:* Access the evolution controller for the classified evolution-indicating events that occurred for the link target since the last update of the age;
- Step 2:* Look up the rating of the occurred classified events and compute the new link age age_{New} incrementing the current age by the sum of those ratings;
- Step 3:* Access the strategy manager and look up the aging strategy identified by the optional attribute *strategy* or by the link type;
- Step 4:* Determine if the threshold value of a new age level has been reached by age_{New} for the chosen strategy. In this case update the age level;
- Step 5:* Enter the current date as the date of the last age update;
- Step 6:* Repeat steps 1 - 5 for all aging links in the considered information object.

The Strategy Manager. For the strategy management a tool has been implemented in Java which enables the definition of new aging strategies (e.g. for new link types) and the update of existing ones. Supported updates are changing threshold values, adding or deleting age levels, etc. Figure 4 shows the user interface for mapping threshold values to age levels. A further functionality of this tool is the rating of evolution indicating events.

Handling Evolution Indicating Events. In order to test our aging server, we constructed an environment for the generation and handling of evolution-indicating events. This environment consists of an XML-based event controller and an XML editor with a simple version management. For each considered target information object, the controller stores the occurred evolution-indicating events in a separate XML file and, on request, passes them to the *Ager*.

Currently only two classes of evolution-indicating events are supported, namely minor and major changes. The author classifies the changes manually. The

¹ The incorporation of a JavaScript function into the stylesheet enables the inspection of the link type and age level of each link in an extra window (see Fig.1).

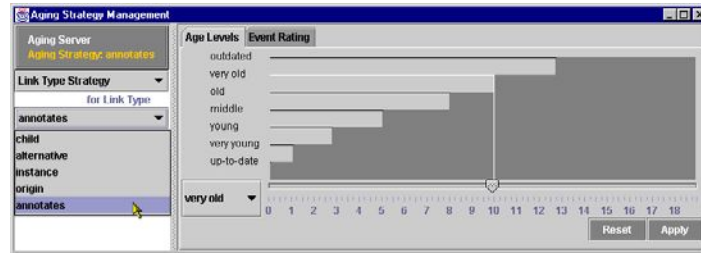


Fig. 4. The Management of Aging Strategies

event controller is informed about the new version generation and its classification.

5 Related Work

Our work on link aging has been inspired by research in different areas: The definition of expiry dates as it is specified in the HTTP [1] builds upon a simple notion of time-driven aging: A piece of information, here an HTML document, is only valid for a certain amount of time and loses validity when the expiry date is reached. In contrast to our approach, aging applies to documents, not to links. An expiry date may be considered as a simple form of a topicality lifeline (see Sect.3.3).

As link aging, link checking for Web sites is concerned with the effects of target object evolution on links. It addresses a very basic but also drastic kind of evolution: The referenced information object has been deleted or moved away. Link checking can be automated; different tools for this purpose exist (e.g. [6]).

Enriching links with additional properties, we conform with a number of other projects. In [15] links are equipped with probabilities which, combined with other facts, improve information retrieval effectivity within hypertext systems. We hold the view that link probabilities can be combined with link ages, as the age of a link may also influence the probability of link relevance.

Evolution awareness is an important topic in CSCW research and system development. Methods for its improvement range from a very close coupling of the cooperation partners as it is implemented by WYSIWIS editors [21] over more relaxed methods as proposed in [7] to asynchronous notification in separate information spaces [14]. Since aging supports close as well as loose cooperation it could be effectively combined with other awareness methods.

6 Conclusions and Further Research

In this paper we present ongoing work on link aging as a non-intrusive mechanism to improve evolution awareness in cooperative work with linked information. Customization through aging strategies enables an adaptation of the aging process to different link types and project situations. The generalization of link aging allows for a smooth integration of time- and change-based aging, covering an extensible set of aging process variants. This ensures the applicability to domains

with different evolution characteristics. First results with the *EvEnAge* prototype show that it can be easily integrated into existing link frameworks.

As a next step we will integrate a differencing engine, developed at our department [20]. This engine is based on a classification and rating scheme for document change operations which can be exploited for rating evolution-indicating events.

Our approach to link aging can be improved by user interaction. Therefore, we will integrate an option for interactively changing the age level, e.g. resetting it after validity has been checked by user intervention. The choice of age level thresholds and the rating of evolution-indicating events is crucial for the correct course of the aging process. With this in mind, we will experiment with different methods to facilitate their adequate fixing, e.g. the inclusion of user feedback, the statistical evaluation of value choices, and the collection of threshold profiles.

Within our system, we represent age levels as discrete time intervals. Modeling age by step functions can only be considered as a first approximation since aging is a continuous process. Thus, for our further research we will model age levels as fuzzy sets (see e.g. [23]). The resulting vagueness of age levels can be exploited for proximity queries and more intuitive graphical user interfaces.

Acknowledgments. The research described in this paper was partly supported by the HSPIII Project WEL and the DFG Project KOLIBRI (DFG Schm 450/7-1 and DFG MA 2005/1-2). Thanks are due to Helen Pert and Jack Koftikian for comments on the manuscript.

References

- [1] T. Berners-Lee, R. Fielding, and H. Frystyk. Hypertext Transfer Protocol – HTTP/1.0. <http://www.w3.org/Protocols/HTTP/1.0/spec>, February 1996.
- [2] S. Chawathe and H. Garcia-Molina. Meaningful Change Detection in Structured Data. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 26–37, New York, May 1997.
- [3] J. Clark. XSL Transformations (XSLT) Version 1.0, <http://www.w3.org/TR/xslt>, November 1999.
- [4] S. DeRose, E. Maler, D. Orchard, and B. Trafford. XML Linking Language (XLink) - W3C Working Draft 21-February-2000, <http://www.w3.org/TR/xlink>.
- [5] P. Dourish and V. Bellotti. Awareness and Coordination in Shared Workspaces. In *Proceedings of the ACM CSCW'92, Toronto, Canada*, pages 107–114. ACM-Press, November 1992.
- [6] R.T. Fielding. Maintaining Distributed Hypertext Infostructures: Welcome to MOMspider's Web. In *First International World-Wide Web Conference (WWW94) in Geneva, Switzerland*. Elsevier Science, May 1994.
- [7] S. Greenberg, C. Gutwin, and A. Cockburn. Using Distortion-Oriented Displays to Support Workspace Awareness. In *Proceedings of the HCI'96 Conference, London*, pages 299–314. Springer-Verlag, August 1996.
- [8] K.M. Hansen, C. Yndigegn, and K. Grønbaek. Dynamic Use of Digital Library Material - Supporting Users with Typed Links in Open Hypermedia. In *Proceedings of the ECDL'99 Conference, Paris, France*, page 254 ff. Springer-Verlag, September 1999.
- [9] IBM alpha Works. XML Parser for Java. <http://www.alphaWorks.ibm.com>.

- [10] F. Matthes and U. Steffens. PIA - A Generic Model and System for Interactive Product and Service Catalogs. In *Proceedings of the ECDL'99 Conference, Paris, France*, pages 403–422. Springer-Verlag, September 1999.
- [11] J. Munson and P. Dewan. A Flexible Object Merging Framework. In *Proceedings of the ACM CSCW'94 Conference, Chapel Hill, NC*, pages 231 – 242. ACM-Press, October 1994.
- [12] NetMind. Mind-It Notification Service. <http://www.netmind.com/>.
- [13] C. Niederée, U. Steffens, J.W. Schmidt, and F. Matthes. Improving Evolution Awareness through Aging Links. Technical report, Technical University Hamburg-Harburg, August 2000. <http://www.sts.tu-harburg.de/papers/STS-Rep-2000.html>.
- [14] U. Rauschenbach. Supporting Awareness in Shared Workspaces Using Relevance-dependent Event Notifications. In *Proceedings of the CVE'96 Workshop, Nottingham*, September 1996.
- [15] T. Rölleke and M. Blömer. Probabilistic Logical Information Retrieval for Content, Hypertext, and Database Querying. In *Hypertext - Information Retrieval - Multimedia, HIM'97*, pages 147–160. Universitätsverlag Konstanz, 1997.
- [16] J.W. Schmidt, G. Schröder, C. Niederée, and F. Matthes. Linguistic and Architectural Requirements for Personalized Digital Libraries. *International Journal of Digital Libraries*, 1(1), 1997.
- [17] K. Shafer, S. Weibel, and J. Fausey. Introduction to Persistent Uniform Resource Locators. <http://purl.oclc.org>, 1996.
- [18] Software Systems Institute, Technical University Hamburg-Harburg. Homepage of the KOLIBRI project. <http://www.sts.tu-harburg.de/projects/Kolibri>.
- [19] Software Systems Institute, Technical University Hamburg-Harburg. Homepage of the Warburg Electronic Library project (WEL). <http://www.sts.tu-harburg.de/projects/WEL>.
- [20] C. Spreen. Change Management für Metadaten in Digitalen Bibliotheken. Master's Thesis, February 2000. <http://www.sts.tu-harburg.de/papers/2000/Spre00>.
- [21] M. Stefik, D.G. Bobrow, G. Foster, S. Lanning, and D.G. Tatar. WYSIWIS Revised: Early experiences with Multiuser Interfaces. *ACM Transactions on Office Information Systems*, 5(2):147–167, 1987.
- [22] R. H. Trigg and M. Weiser. TEXTNET: A Network-Based Approach to Text Handling. *ACM Transactions on Office Information Systems*, 4(1):1–23, 1983.
- [23] L. A. Zadeh. A Theory of Approximate Reasoning. *Machine Intelligence*, 9:149–194, 1979.

Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints

Linda L. Hill

Alexandria Digital Library Project
University of California, Santa Barbara
lhill@alexandria.ucsb.edu

Abstract. The core elements of a digital gazetteer are the placename itself, the type of place it labels, and a geographic footprint representing its location and possibly its extent. Such gazetteer data is an important component of *indirect geographic referencing* through placenames. Based on the gazetteer development work of the Alexandria Digital Library, this paper presents the nature of placenames, and the process of assigning categories to places based on the words in the placenames and other information, and discusses the nature of georeferencing places with geographic footprints.

1 Introduction

This paper concerns digital gazetteers, which can be defined as *geospatial dictionaries of geographic names* with the core components of

- a name (could have variant names also)
- a location (coordinates representing a point, line, or areal location)
- a type (selected from a type scheme of categories for places/features)

Digital gazetteers have new roles to play in information architectures for geolocal access to information and data, where the placenames we use in everyday discourse to identify a location can be used to get (from computer systems) directions for driving there or finding information about an area. Structures for the interoperability of digital gazetteers hold promise for networked access to diverse sets of gazetteer data to maximize availability of global, local, and specialized place/feature descriptions and their spatial locations.

First, background information on the development of digital gazetteers and digital gazetteer information exchange (DGIE) will be given as a foundation for what follows. The next sections focus on the three required elements of a digital gazetteer: the nature of the placenames themselves; the process of assigning type categories to places/features; and a discussion of some of the issues of representing feature locations with geospatial *footprints* (e.g., latitude and longitude coordinates). This paper builds on previous papers describing the Alexandria Digital Library gazetteer development [1-3] and on online resources that are referenced herein.

2 Current Status of Digital Gazetteers

The U.S. Board on Geographic Names (BGN) supports the provision of two gazetteers; both of them are accessible through web services. They are: the U.S. Geological Survey's (USGS) Geographic Names Information System (GNIS) <http://www-nmd.usgs.gov/www/gnis/> for domestic features, and the National Imagery and Mapping Agency's (NIMA) Geonames Server <http://www.nima.mil/gns/html/>. There are other well-established gazetteer web services, for example, the Canadian Geographic Names Service <http://GeoNames.NRCan.gc.ca/english/> and a growing number of new ones. There are many other placename services that do not call themselves *gazetteers* but serve similar purposes. These include *yellow page* services for commercial locations, and placename access to online mapping and routing services and digital atlases. There are web sites that provide pointers to online gazetteers; a good starting point for such an exploration is the USGS GNIS web site.

The Getty Thesaurus of Geographic Names (TGN) http://shiva.pub.getty.edu/tgn_browser/ is a gazetteer of a different nature. It is one of three "structured vocabularies that can be used to improve access to information about art, architecture, and material culture.... Linked to the record for the place are names, a 'parent' or position in the hierarchy, other relationships, geographic coordinates, notes, sources for the data, and 'place types,' which are terms describing the role of the place (e.g., 'inhabited place' and 'state capital')." Names are in various languages and syntactic forms. One name is flagged as the preferred name. TGN does not require that records include geographic coordinates, but rather it contains a mix of entries with and without geographic footprints.

A characteristic of these gazetteer and gazetteer-like services¹ is that there is *no* standardization of any aspect of the record format, the content, or the service interfaces. Data cannot be easily shared among these resources. Standardization for official gazetteer efforts has been focused on establishing authoritative name forms for non-commercial places only. Official gazetteer agencies have not been involved with establishing authoritative boundary representations for the extents of features; their purposes have been satisfactorily served with point locations that disambiguate one place from another. Each gazetteer and gazetteer-like service also has created its own set of categories for placenames and its own record format and description details.

Implementations of digital libraries and online georeferenced collections and services have now created new roles for digital gazetteers. Placenames that are both geospatially referenced and categorized with a common scheme are the key to *indirect geospatial referencing*, meaning the use of a placename to identify a geographic location. For example,

- Using *Lisboa* or *Lisbon* instead of a coordinate expression (e.g., a point location of 9°7'59"W and 38°43'0"N) to find geospatial information about that particular location on our planet. A query such as "What satellite images

¹ *Gazetteer-like* services and collections include the placename-location data inherent in GIS databases, the commercial yellow-page listings of named locations and spatial references, the Census address-location files, and all listings of named features that include spatial location representation.

cover the area of *Lisbon*?" can be translated into a query with the latitude and longitude coordinates for *Lisbon* by using a digital gazetteer. This query can be sent to a collection of such objects where the only georeferencing is by coordinates or other geospatial referencing system - not by placenames.

- Using *Santa Barbara* to establish a point location on Earth so that sky maps and satellite locations can be oriented to the particular user's location. This method is used in the Heavens Above website from German Space Operations Center <http://www.heavens-above.com/>.
- Using placenames to identify point locations for navigation purposes, such as the GO2 system of wireless navigation services where users will be able to enter a short name identifier for, say McDonald's, and find the nearest McDonald's locations to the user's current location www.go2online.com.

These new roles for digital gazetteers result in new requirements for gazetteer data. Some applications (e.g., digital libraries such as the Alexandria Digital Library (ADL)) require geographic footprints giving spatial extent, not just point locations. Some applications require a mix of commercial and public places, of natural and manmade features, and of local and global locations. Some are based on street addresses and postal codes. Knowledge about historical places is of great interest to cultural and humanities activities and to genealogists. Information about geographic features and their names and characteristics is held in many ways by many organizations and individuals. What we need are ways to make this information shareable through search and retrieval protocols designed for gazetteer data, exchangeable record formats, defined elements of description, and common schemes for categories. Without a framework for generating and gathering and accessing gazetteer data, we will fall short in the areas of comprehensive placename knowledge and richness of representation and description.

3 Alexandria Digital Library Gazetteer

The ADL Project at the University of California, Santa Barbara, has included gazetteer development in its research from the beginning of its NSF-funded project in 1994. Details of this development are described elsewhere [1, 3] and current information can be found on the ADL website www.alexandria.ucsb.edu.

This gazetteer was built by loading available sets of gazetteer data. The datasets loaded into the ADL Gazetteer include those of the two U.S. federal agencies (U.S. National Imagery and Mapping Agency (NIMA) and the U.S. Geological Survey) plus U.S. topographic map areas, California earthquake epicenters, volcanoes from the Smithsonian Global Volcanism Program, and additional sets of bounding boxes for administrative areas.

The ADL Gazetteer data is described according to the ADL Gazetteer Content Standard and the ADL Feature Type Thesaurus. The ADL Gazetteer is available both as a web service www.alexandria.ucsb.edu/gazetteer/gazserver and as a component of the Alexandria Digital Library (the operational ADL). The operational ADL includes a catalog of georeferenced holdings of the Map & Imagery Laboratory (UCSB Davidson Library) with more than 2 million items as well as the ADL gazetteer with over 4 million entries; access to the operational ADL through either UCSB or through

the California Digital Library is restricted to connections from the UC campuses only (as of April 2000).

This paper does not go into detail about the ADL Gazetteer Content Standard. Please consult previous papers and the ADL Gazetteer Development web page www.alexandria.ucsb.edu/gazetteer for the Content Standard itself, its relational database implementation, and the XML DTD version of it. There is also a link to information about the two-day workshop on Digital Gazetteer Information Exchange (DGIE), sponsored by the National Science Foundation (NSF) and hosted by the Smithsonian Institution, which was held in October 1999.

4 Nature of Placenames and Categories of Places

ADL gained experience with the typing of placenames through the merging of the two U.S. federal gazetteers and the development of a Feature Type Thesaurus that draws on the terminology of existing typing schemes. An inventory of existing classification schemes for features identified the following representative schemes:

Table 1. Pointers to representative type schemes in current online gazetteers²

Organization	URL
U.S. Geological Survey	http://mapping.usgs.gov/www/gnis/features.html
U.S. NIMA	http://www.nima.mil/gns/html/FD_Cross_Ref.html
Australian Geographic Names	http://www.environment.gov.au/database/MAN200R.html
Canadian Geographical Names	http://GeoNames.NRCan.gc.ca/english/cgndb.html
Australia's National Mapping Agency	http://www.auslig.gov.au/mapping/names/names.htm
Getty Thesaurus of Geographic Names	http://shiva.pub.getty.edu/tgn_browser/about_tgn.html "Place types are indexing terms chosen from the structured vocabulary of the AAT"
ADL	http://www.alexandria.ucsb.edu/gazetteer/FeatureTypes/index.htm

There is no shared typing scheme among those listed in Table 1. The ADL Feature Type Thesaurus (FTT) was developed with the idea that it could provide a bridge among the various typing terminologies. To the greatest extent possible, all of the terminology discovered in the other schemes was added to the FTT as lead-in vocabulary, pointing to the preferred term for that concept in the FTT if it is not a valid term itself. Frequently occurring *surnames* in placenames were also added to the FTT (*surnames* are described below). The resulting thesaurus has close to a 5:1 ratio of lead-in vocabulary to preferred terms (209 preferred terms and 978 lead-in terms).

When the U.S. federal gazetteer data was loaded into the ADL Gazetteer, conversion programs were written in Perl that assigned an ADL Feature Type to each

² Classification schemes for commercial activities, such as the U.S. Standard Industrial Classification (SIC) system and the North American Industrial Classification System (NAICS), are not listed in the table because they are designed for a different purpose. However, they are also a type of classification for sites that can be geo-located.

entry based on (1) the content of the placenames and (2) the classification of the originating agency. The most difficult set to convert was the GNIS from the U.S. Geological Survey. The difficulty was caused primarily by their use of broad categories that lumped discrete types together. For example, in the GNIS PARK category, ADL analyzed the placenames and found the following types of places, based on the placenames themselves: archeological sites, historical sites, hydrographic structures, forests, sports facilities, reserves, monuments, lakes, cemeteries, and museum buildings.

Placenames often contain information about the types of places they are labeling. Consider the examples in Table 2. *Surname* refers to the end position of the placename. In the ADL application, *surnames* can be either single words or two-word phrases. *Keyword* refers to a type indicator that is found in other positions in the placenames, as one entry in Table 2 illustrates; the keyword “correctional” appears in combination with a variety of other words in placenames and the word itself is a reliable indicator of the type “correctional facilities.” The use of *keyword* terminology must be carefully done to avoid mis-assigning types to placenames. A general rule for such *keyword* assignments is that if the *surname* (the end word or words) is recognized from the list of surnames, then its assignment has precedence over other *keywords* in the name. For *Railroad Station Café*, the category assignment would be based on *Café*, not on *Railroad* or *Station*.

Table 2. Placename examples where type is represented in the name

Placename	1-Word Surname	2-Word Surname	Keyword
Perth Airport	Airport		
Baldwin County	County		
Admiralty Oil Seep		Oil Seep	
Jar Qudug Gas Field		Gas Field	
Sussex Correctional Institution			Correctional

Some placenames don’t carry enough information within the name to distinguish a type category. For these, the classification must be provided by an outside source; for example, classifications assigned by an official gazetteer agency, from local knowledge, or from associated description. Note that many common placenames do not include their category, including countries (“Mexico,” not “Country of Mexico”), U.S. states (“California,” not “State of California”), and many cities (“Boston,” not “City of Boston”).

Some terms used in placenames are applied to various types of locations and therefore cannot be used alone to assign a type. The surname ‘Field’ is used as an example in Table 3.

Add to this complexity the tendency of humans to be cute when they name shopping centers, housing developments, and so forth, and the use of historical names for buildings that are now used for other purposes, and you have a set of names that will be mis-typed if the names alone are used as a source of evidence for choosing a type category. Reliable manual typing is the best solution for these cases - and for all cases.

Table 3. Examples of ambiguous type term in placenames

Placename	Classification
Kindley Field	airports
Bilate River Field	volcanoes
Jervis Field	agricultural sites
Lance Field	populated places
Eastern Fields	bars (physiographic)
Titas Field	oil fields

Our task was to convert the GNIS gazetteer to our ADL Gazetteer. To do this, programs were written in Perl to analyze the GNIS data in sets corresponding to the GNIS categories. The programs are of this form:

- Step 1. Normalize the placenames by removing parenthetical phrases, numbers, directions (e.g. “North”), and trailing character strings from the ends of the names; normalize spacing and punctuation; and convert to lower case.
- Step 2. Extract {two-word, one-word} surnames from sets of GNIS data, one set for each of their categories; keep surnames that meet a frequency of occurrence threshold; search the placenames for the occurrence of these surnames at the beginning and anywhere within the placenames.
- Step 3. Compare these surnames to the ADL FTT Thesaurus; extract matches to valid and lead-in type terms.
- Step 4. Create reports for two-word and one-word surnames, including the matching FTT terms that were found.
- Step 5. Sort lists by frequency of occurrence of surnames and manually review.
 - a. Accept automatic assignment of term(s) from FTT or change them
 - b. Modify FTT as necessary
 - c. Add type assignments for terms that did not match to FTT
 - d. Set treatment of surnames to “keyword” if the surnames occur in other positions in the names with the same meaning.

Geographic Name *
Name * [the primary name for feature in a particular the gazetteer application]
Name Source
Etymology
Language (default is English)
Pronunciation
Transliteration Scheme Used
Character Set (default is ASCII)
Current / Historical Note * (default is Current)
Beginning Date
Ending Date
Time Period Note
Source Mnemonic
Entry Date
Variant Geographic Name (R)
 [same attributes as for primary name]

*** Required (R) Repeatable**

Fig. 2. Attributes for placenames in ADL Gazetteer Content Standard

The actual conversion process included monitoring to catch as many anomalies as possible while still processing large sets of data. Where type classification could not be reliably based on placename analysis, typing was based on a default mapping from

the GNIS category to the ADL type. Where the GNIS category was too broad for a default mapping (e.g., their OTHER category) and the placenames were ambiguous, ADL noted the entries as *uncategorized*; .5% of the ADL Gazetteer entries are uncategorized.

We know that this automated process of assigning categories to places is not totally accurate and that mis-typing has occurred despite our best efforts. There is no substitute for direct knowledge of the places when assigning categories. We would like to be able to enlist the help of groups and individuals with local knowledge to review and revise the category assignments as necessary, as well as to contribute additional data and information. To this end, we have developed an XML version of the ADL Gazetteer Content Standard and the processes to export and import data in this format.

We have plans to mine the ADL Gazetteer of over 4 million entries using the surname extraction programs to produce a frequency distribution of surnames per category and categories per surname. This data will be interesting on its own, but it can also be used to provide automated support for assigning categories to new gazetteer entries and to support query formation in user interfaces.

All of these examples are in English and reflect Anglo-Saxon naming conventions, English-language characteristics of word construction, and perhaps Anglo-Saxon sets of categories. Extension to other languages will require customization to accommodate naming conventions and other character sets. There is an interesting possibility that a gazetteer that includes placenames in multiple languages would be able to bridge language differences. This could be done by analyzing the co-occurrence of non-English language keywords in the placenames to type categories based on the English language placenames. The ADL Gazetteer Content Standard and some existing gazetteer datasets (e.g., the NIMA gazetteer) include a language attribute for placenames (Figure 2) to support this type of category assignment across languages.

Street addresses and coding schemes such as postal codes and the U.S. Federal Information Processing System (FIPS) codes have not been included in this discussion because assigning categories to them is not complicated. The merging of retail sites with the types of place/features typically found in gazetteers will necessitate an analysis of the cross-over of commercial classification systems, such as the U.S. Standard Industrial Classification (SIC) system and the North American Industrial Classification System (NAICS), to gazetteer classification systems like the ADL Feature Type Thesaurus in the future.

5 Georeferencing Places

A geographic *footprint* representing the location of a named place is the other requirement of a digital gazetteer, as defined here. This footprint, in latitude and longitude coordinates, can be one of the following (Table 4):

Table 4. Types of geospatial representation

Type of Representation	Description
point	single pair of latitude & longitude coordinates
bounding box	double pair of coordinates representing the maximum and minimum of latitude and longitude extent
line	set of points that do not enclose a space
polygon	set of points that do enclose a space
grid representation	grid references to a location according to an identified grid referencing scheme

A place can have multiple footprint representations:

- of different types: a point, bounding box, and a polygon, for example
- from different sources
- for different time periods (for example, the extents of cities change through time)
- suitable for varying resolutions (e.g., more detailed vs. more generalized)

Such a set of footprints bound to a named object (a *placename location*) would support a number of useful operations:

- viewing of footprint changes through time³
- comparison of footprints from different sources
- transformations between footprint representations

Placenames are inherently imprecise. Goodchild described this well in his presentation to the Digital Gazetteer Information Exchange workshop:

We work in a world with words like "hot" and "cold", which have very vague meanings, and our vernacular geography is also vague. I come from Santa Barbara, and I don't worry about the precise definition of the footprint of Santa Barbara. In fact, I don't come from the municipality of Santa Barbara, if you wanted to be fussy

Detailed Spatial Geometry Representation (R)

[set of points; dependent on system capabilities and requirements; can represent set of non-contiguous areas]

Detailed Spatial Geometry Representation

* {point, bounding box, linear, complex object}

Number of Points *

Points Order *

(Longitude, Latitude) *

Current / Historical Note *

Beginning Date

Ending Date

Time Period Note

Measurement Date, Beginning Date

Measurement Date, Ending Date

Method of Measurement

Accuracy of Measurement

Source Mnemonic

Entry Date

* **required element** (R) repeatable

Fig. 3. Specification of location in ADL Gazetteer Content Standard

³ There has been some discussion about whether or not the gazetteer structure would be appropriate for named events like hurricanes where the temporal dimension is the "run time" of the event and the footprints are rapidly changing.

about it, but you know what I mean when I say that I come from Santa Barbara. On the other hand, our GIS's are exact and "scientific"; they are in a very different, precise world. In many ways a gazetteer provides the linkage between those two worlds. It provides the linkage between the vernacular terms we use to talk about the world and the precise coordinate systems we use when we need to be exact.... [I]f one wants to combine vague and precise, then one has to deal explicitly with accuracy. The accuracy of a placename specification has to be an explicit part of the future gazetteer. We need to know how vague "vague" is. [4]

Each footprint used to describe a place is to some degree an approximation, which should be documented from the supplier to the end user. Often we depend on the nature of points and bounding boxes to convey a sense of approximation without further documentation. *Best practice*, however, would include documenting all footprints as specified by the ADL Gazetteer Content Standard (Figure 3) or in accord with the metadata standards for geospatial data [5].

Documentation of accuracy can be used to guide the appropriate use/interpretation of placenames and footprints, which is highly dependent on scale, ranging from local to global contexts. To quote Goodchild further:

When location is specified, scale often provides a great deal of context. So when I say "Santa Barbara" at a global scale, you attach a different meaning to that than if I use the term at a local scale. "Santa Barbara" at a local scale probably means the municipal boundaries; "Santa Barbara" at a global scale probably means something much more vague - some blot on the map of Southern California. [4]

Sources of georeferencing include existing gazetteers, GIS datasets, numerous specialized listings, cartographic products, environmental analysis documents, research publications, measurements with Global Positioning Systems (GPS) units, and more. The problem isn't that these footprints don't exist, but rather that most of them are not easily accessible as gazetteer data. The key to making them accessible is a system of consistent naming (labeling) linked to the various footprint representations. That is, geographic names (location names) become primary keys to the spatial representations, and the geographic names themselves are *well known* in the sense that they can be referenced to an authoritative gazetteer source. Authoritative sources should (but they don't always) eliminate duplicate entries for the same places to minimize confusion.

Satisficing is a criterion that needs to be applied to the choice of gazetteer footprint representation for various purposes.

Satisficing has been defined by Simon (1979) as a psychological form of the law of diminishing returns – we do not seek optimal solutions to certain problems because the costs are too high, so we settle for solutions that are satisfactory given the cost. [6]

Intuitively, we know that a point location is suitable for locating a person in space, for example, and for representing point-like entities like wells or monuments, or for

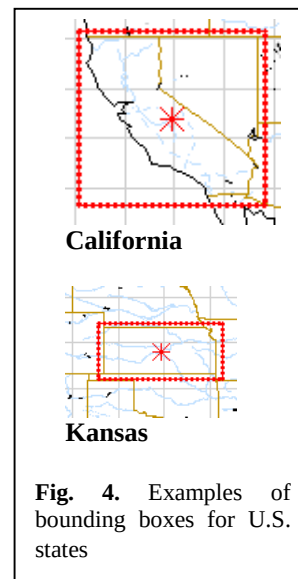


Fig. 4. Examples of bounding boxes for U.S. states

representing large entities from a great distance (scale dependencies). Points for linear features like streams are satisfactory for the limited purpose of disambiguating one stream from another, but are not satisfactory for representing the extent of the feature or locating neighboring features.

We have less intuitive understanding of whether bounding boxes are *satisficing* specifications of locations in information retrieval systems. We can be assured that the extent of a bounding box encloses the extent of the feature but we cannot anticipate the similarity of the bounding box to the shape and coverage of the feature. This similarity depends on the characteristics of the shape of the feature. Figure 4 shows two examples of bounding boxes for U.S. states. It is obvious that the bounding box for Kansas is a better fit than the bounding box for California. We can measure this similarity for each instance but we don't have a formal way to specify in general the relationship of shape to bounding box and thus what we can expect from a general application of bounding box representations to geographic features. We have to estimate sufficiency of representation for bounding boxes by how they perform as a set for a particular purpose.

For spatial matching operations in information retrieval systems, bounding boxes have an efficiency advantage over more detailed polygons; spatial *overlap*, *contains*, and *is-contained-in* operations can be efficiently performed on bounding boxes. In what circumstances is it satisfactory then to rely on bounding box generalizations in spatial information retrieval systems? If generalized polygons that more closely fit actual feature shapes are used instead of bounding boxes, what is the best balance of point density to efficiency and sufficiency for information retrieval? Hill [7] finds evidence to support the premise that ranking retrieval sets by spatial similarity to a reference region (e.g., a query region) "is relatively insensitive to differences in the exact shape and size of the [footprints]." This premise needs to be tested with more formal methods for spatially enabled information retrieval systems and other applications.

6 Conclusions

Three elements of digital gazetteer data have been the focus of this paper: the placenames themselves, the process of assigning type categories to places/features, and the representation of geospatial location. These three elements form the core descriptive requirements for digital gazetteers; optional elements include descriptions, links to related gazetteer entries, associated data, and links to online resources about the place/feature. Many communities have a stake in building and accessing digital gazetteers. The key to increasing their usefulness is interoperability. The first step is the recognition of the importance of *indirect geographic referencing* through placenames to digital earth information system applications. The second step consists of working together to establish the standards necessary to share gazetteer data across numerous distributed activities. These standards include applying techniques of controlled vocabularies (i.e., thesauri and domain lists) to bring consistency to the terms used to designate categories, and making such schemes themselves interoperable. This is a difficult but not impossible task and will be aided by automated processes of placename analysis such as those discussed here.

Acknowledgements. The Alexandria Digital Library gazetteer development work has been financially supported by NASA Research Grant NAG5-8542 (ESDIS Prototype, Goddard Space Flight Center), by National Science Foundation Grant EIA99-78209 (a Digital Government grant for the DGIE workshop), and by the NSF Grant IR94-11330 (Digital Libraries I). UCSB graduate student Doug Tallman, Bren School of Environmental Science and Management, programmed the surname extraction programs. Ajay Kochhar, UCSB graduate student in computer science, has helped in converting sets of data and loading them into the ADL Gazetteer. Michael Goodchild contributed editing comments to a draft version.

References

1. Hill, L.L., Frew, J., and Zheng, Q.: Geographic Names: The Implementation of a Gazetteer in a Georeferenced Digital Library. D-Lib, January (1999). <http://www.dlib.org/dlib/january99/hill/01hill.html>.
2. Hill, L.L.: Gazetteer and Collection-level Metadata Developments. In Kaser, R.T., Kaser, V.C. (eds.): *Metadiversity. The Grand Challenge for Biodiversity Information Management through Metadata. The Call for Community*. Proceedings of the Symposium sponsored by the U.S. Geological Survey Biological Resources Div. & the National Federation of Abstracting & Information Services. NFAIS (1999) 141-145.
3. Hill, L.L. and Zheng, Q.: Indirect Geospatial Referencing through Place Names in the Digital Library: Alexandria Digital Library Experience with Developing and Implementing Gazetteers. In *Knowledge: Creation, Organization and Use*. Proceedings of the 62nd Annual Meeting of the American Society for Information Science, Washington, D.C., Oct. 31-Nov. 4, 1999. Information Today, Medford, NJ (1999) 57-69. www.alexandria.ucsb.edu/~lhill/paper_papers/ASIS99_confpaper2_final.pdf.
4. Goodchild, M.F. The Future of the Gazetteer. Presented at the Digital Gazetteer Information Exchange Workshop, Oct 13-14, 1999. Transcribed and edited from audiotape. http://www.alexandria.ucsb.edu/~lhill/dgie/DGIE_website/Perspectives/brandt.htm.
5. U.S. Federal Geographic Data Committee. Content Standard for Digital Geospatial Metadata. [Accessed 4/21, 2000]. <http://fgdc.er.usgs.gov/metadata/contstan.html>.
6. Marchionini, G. *Information Seeking in Electronic Environments*. Cambridge Series on Human-Computer Interaction. Vol. 9. Cambridge University Press (1995).
7. Hill, L.L. *Access to Geographic Concepts in Online Bibliographic Files: Practices and the Potential of a Graphic Interface*. Ph.D. Dissertation. University of Pittsburgh, School of Library and Information Science (1990) 113.

The Application of an Event-Aware Metadata Model to an Online Oral History Archive

Jane Hunter¹ and Darren James²

¹CITEC, jane@dstc.edu.au

²DSTC Pty. Ltd., d.james@dstc.edu.au

Abstract. In this paper we test the ABC event-aware metadata model, developed within the Harmony project, by applying it to a complex multimedia oral history archive. Based on a metadata schema, generated using the ABC model, we developed indexing tools, a database and a search and browse Web interface, for an oral history collection consisting of audio tapes and posters generated from a series of interviews and photographs. The objective was to build a test-bed to test and refine the ABC model and also to demonstrate that use of the model will ensure consistent, well-structured, unambiguous metadata descriptions for complex multimedia collections. Such descriptions will hopefully lead to improved fine-grained resource discovery, interoperability between different metadata schemes and explicit tracking of intellectual property rights.

1 Introduction

The future will lead to many more compound multimedia documents on the web which combine text, image, audio and video in rich complex structured documents in which temporal, spatial, structural and semantic relationships exist between the components. The problems associated with indexing, archiving, searching, browsing and retrieving these kinds of structured dynamic documents are infinitely more complex than the resource discovery of simple atomic textual documents. In addition, the intellectual property rights associated with the components which make up such documents are frequently enmeshed within a complex and dynamic web which is extremely difficult to manage.

The Dublin Core [1], bibliographic [2], museums [3], archival [4] and MPEG-7 [5] communities are developing international standards for describing textual, physical and audiovisual resources to enable their resource discovery over the Internet. The INDECS organisation [6] is developing metadata standards to support network commerce in intellectual property. Managing the new heterogeneous multimedia resources which are being developed and traded via the Internet is going to require interoperability between these disparate standards [7].

The Harmony International Digital Library project [8] is investigating models and tools, which will facilitate machine-understandable communication between the con-

tent description domains (Dublin Core, MPEG-7, IFLA, CIDOC) and the e-commerce and rights management domain (INDECS).

The collaborators are developing a common data model and metadata vocabulary, which will hopefully provide a stable underlying infrastructure for the exploitation of digital multimedia content over the Internet.

In this paper we apply an event-aware metadata model [9], which has been developed within the Harmony project, to provide a clear understanding of the individual resources, events, contributions and relationships which constitute a complex collection of digital multimedia resources. By applying the proposed model, we hope to generate consistent, well-structured, unambiguous and modular metadata descriptions which will lead to improved fine-grained resource discovery, easier mapping between different metadata schemes and the explicit tracking of intellectual property rights.

In order to test the model and the quality of the derived metadata, we built an online digital audio (MPEG-2 Layer3 (MP3)[10]) and image archive, together with indexing tools and a search and browse interface, for an oral history project. The *From Lunchroom to Boardroom* project [11] combines oral, written and visual art techniques to highlight the achievements of women in the Labor movement in Queensland from the 1930's to the 1960's. The collection consists of posters and audio tapes generated from a series of interviews and photographs. It provides an ideal multimedia testbed for applying and testing the proposed ABC metadata model. In addition, the development of a digital online version will greatly increase the accessibility and dissemination of this valuable historical collection.

Based on a metadata schema which was derived by applying the proposed model to this collection, we built a multi-level metadata editor/generator application, MEGGIE. This Java application includes a computer-assisted audio file transcriber which enables the input of transcripts and their alignment with the corresponding digital speech files. The structured and validated metadata descriptions can be output as: ID3 tags [12] embedded in the MP3 files; Dublin Core descriptions encoded in RDF [13]; MPEG-7 descriptions [14] encoded in XML [17]; HTML metatags [15]; or tables in an Oracle database with links to the MP3 files. A web-based search interface was also built which enables users to search and browse the archive via speaker name, keywords or transcript content and to retrieve and play the matching segment of the relevant MP3 file. Figure 1 provides an overview of the system architecture.

The rest of the paper is as follows. In Section 2 we describe the ABC model [9] briefly. Section 3 describes the results of applying the ABC model to this project and the metadata schema which was generated. In Section 4 we describe the MEGGIE metadata/editor generator application which was built to enable the computer-assisted generation of metadata descriptions for a number of different domain-specific vocabularies. Section 5 describes the search, browse and retrieval web interface to the collection. The paper concludes in Section 6 with an evaluation of the ABC event-aware metadata model in the context of this digital multimedia collection and the anticipated future work directions.

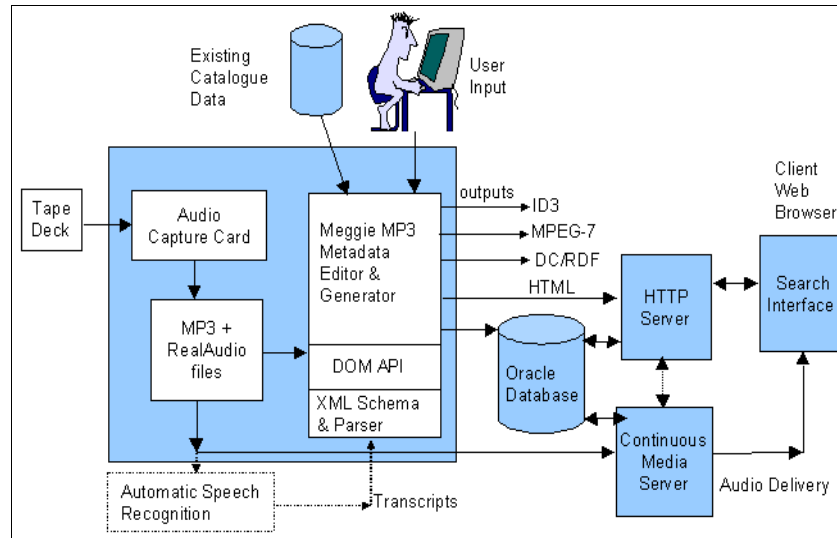


Fig. 1. Overview of the System Architecture

2 The Harmony Event-aware Model for Metadata Interoperability

The Harmony Project [8] is investigating a generic approach towards metadata interoperability and, in particular, its application to multimedia digital libraries. Through the event-aware ABC model and vocabulary, it is attempting to:

- formally define common entities and relationships underlying multiple metadata vocabularies;
- describe them (and their inter-relationships) in a simple logical model;
- provide the framework for extending these common semantics to domain and application-specific metadata vocabularies.

The ABC model [9] defines a set of fundamental classes which provide the building blocks for expression (through sub-classing) of application-specific or domain-specific metadata vocabularies. These base classes were determined by analysing commonalities between different communities' metadata models (including IFLA, CIDOC, INDECS, MPEG-7 and DC):

- Resources
- Events
- Inputs and Outputs
- Acts
- Agents
- Roles
- Context
- Time
- Place
- Event Relations

ABC adopts an event-centric view for modeling the relationship between the various manifestations of a creation. This event-centric view provides semantically clear attachment points for the association of properties amongst the various manifestations and contributors (agents) to the manifestations. In addition ABC provides a multiple views philosophy for metadata modeling and recipes for inter-conversion. When rich information is required, ABC provides the event model. This involves describing the *event* through which that relationship was realized as an object in itself and describing the hidden detail implicit in a simple binary relation. When concise/simple metadata is needed, flatter relations are used.

In the next section we apply this underlying model to the *From Lunchroom to Boardroom* project with the goal being to generate well-structured unambiguous and interoperable metadata descriptions.

3 A Metadata Model for an Online Oral History Archive

3.1 Description of the Application

In order to test the ABC model, we chose a typical example of a complex multimedia archive - an oral history project, which consists of a collection of both analogue and digital resources of multiple media types and involves a complex web of events, contributions and relationships. In this section we outline the background of the project and then we model it diagrammatically using the ABC event-aware metadata model.

The *From Lunchroom to Boardroom* project was first proposed by oral historian, Therese Collie, at the ALP Centenary Celebrations in Barcaldine in Queensland in 1991. Later that year, the project received funding from the Australia Council, the Trades and Labor Council's (TLC) Women's Equal Opportunity Committee and the Public Sector Union of Queensland. During 1991 and 1992, Therese Collie interviewed 16 women of varied backgrounds and living throughout Queensland, who had been involved in the Labor movement between the 1930's and 1960's. The interviews were recorded onto 40x26 minute AGFA professional tapes. These were later transcribed by Sue Pechey and the transcripts were stored as word documents. In addition, visual artist Judith Hewitson, produced four A2 colour posters from photographs and memorabilia collected during the interviewing process. The posters were created using Adobe Photoshop and stored as TIFF images. Each interviewee signed a release form in which they agreed to release all property rights in the tapes and photos to the TLC.

After gaining the approval of the rights owners (the TLC), DSTC digitized the audio tapes into both RealAudio (multi-rate, 56Kbps) and MPEG-2 Layer 3 (MP3) (44.1kHz, mono) formats. The A2 posters were converted to GIF thumbnails and A4 JPEG images. A comparison of RealAudio and MP3 versions of the same 24 minute recordings, demonstrated that, although each MP3 file was approximately 25-30% larger in size, the sound quality was significantly better. MP3 offers sound quality that comes close to the quality of .WAV files but at greatly reduced file size, which means that they download much faster. For this reason, the MP3 file format was selected as the final storage format.

After digitization, the project metadata (for the collection, interviews and recordings) was entered using the MEGGIE application described below. The PEGGIE application, described in [16], was used to index the images associated with the posters. The final step in the indexing process involved the temporal alignment of audio file segments with their corresponding transcript segments. The metadata descriptions were saved to an Oracle database and a Web search interface was developed to enable the retrieval and playback of audio segments based on speaker, word or keyword (subject) searches.

3.2 Applying the Harmony ABC Model

Figure 2 below is a diagrammatic representation of this scenario based on the ABC event-aware metadata model, which in turn adopts the IFLA FRBR model [2] for understanding the lifecycle of intellectual content associated with resources. This model explicitly identifies all of the resources and events, which occurred from the original creation of the concept to the development of the online digital archive.

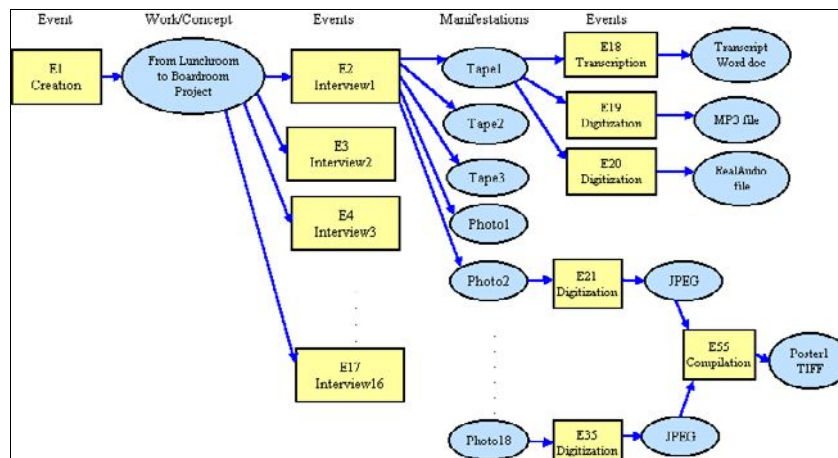


Fig. 2. Event-aware Model of Oral History Project

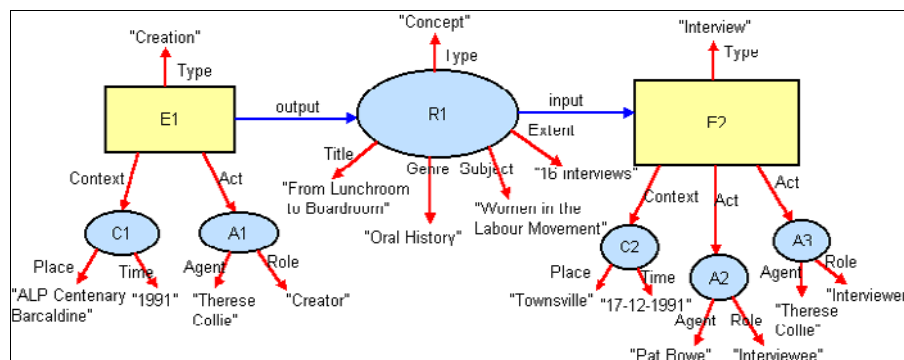


Fig. 3. Modeling the Concept Creation Event and an Interview Event

Figure 3 illustrates the association of properties with events and resources. Under the ABC model, events have inputs, outputs, acts and contexts associated with them. The specific resource attributes will depend on the type of resource and the domain-specific requirements. Figure 3 shows the inputs, outputs and attributes associated with the concept creation event and one of the interview events.

Figure 4 proposes an approach for modeling rights metadata. The output from a rights agreement event (between a rights holder and a rights receiver) is a Right entity. A basic Right entity has a Type (e.g. Permission), a target resource to which it applies, the EventType to which this Right applies and one or more constraints which may be duration-based, fee-based or some other condition.

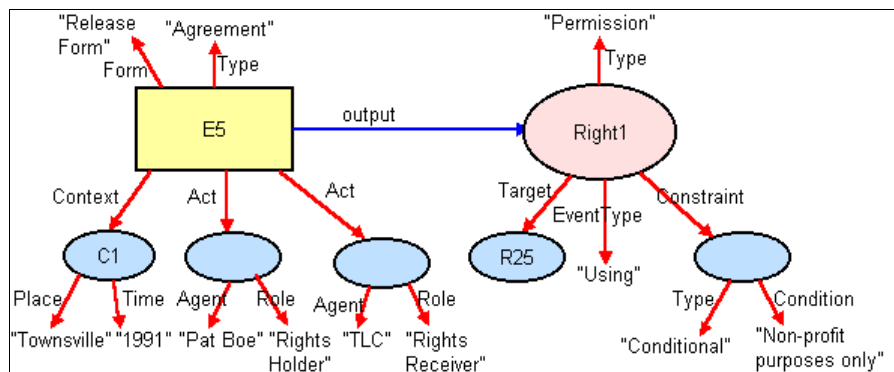


Fig. 4. Modeling a Rights Agreement Event

In this application, we want to be able to link segments of the transcripts to the corresponding audio file segments to enable word or subject (keyword) searching on the transcript and playback of the relevant audio segment. Figure 5 illustrates how the ABC model can be used to model the structural metadata associated with each MP3 file. Using the ABC model it is possible to describe the segments of an audiovisual recording of a real-world event in terms of the sub-events of that real-world event. In this application, we break each interview event into Question-and-Answer sized sub-events. Figure 5 shows the segmentation of the MP3 recordings to reflect this same descriptive granularity.

Using the ABC model, we have been able to completely describe all of the significant information associated with the project in an unambiguous and well-formed manner. The next step is to decide which resources and events need to be described for the particular applications' purposes and then to represent these using the ABC model vocabulary in a schema definition.

For our application we have decided that only metadata for the following four entity types is significant and needs to be recorded:

- The Collection;
- Interviews;
- Digital Recordings;
- Segments.

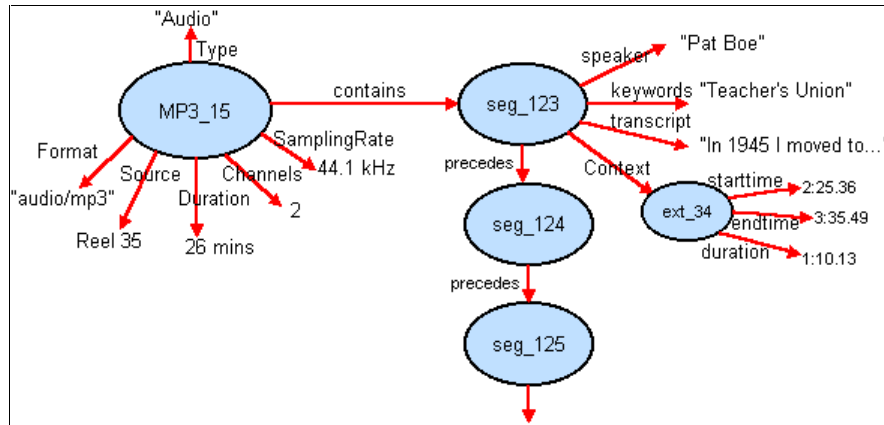


Fig. 5. Modeling MP3 Segment Metadata

3.3 The Schema Representation

There are a number of possible alternatives for defining a schema to constrain and validate descriptions. These include XML DTDs [17], RDF Schema [18] or XML Schema Language [19,20]. We intend using XML Schema language in the near future because of the additional structural, content and datatyping constraints it provides over XML DTDs and RDF. However because of its immaturity and instability at the time of development of this project, we decided to use an XML DTD initially. When we change to an XML Schema to constrain the metadata input, an XML Schema parser will be integrated into the MEGGIE application to perform the description validations. The DOM API [21] generated by the parser will provide a user interface to browse the tree structure of the metadata descriptions and to manipulate their content.

4 The MEGGIE Application

4.1 Inputting the Metadata

MEGGIE (MP3 Metadata Editor and Generator) is a Java application which enables the input and editing of metadata for an oral history collection, based on the underlying XML schema described in Appendix A of the online version of this paper [32]. The application provides input forms for the Collection-level, Interview-level and Recording-level metadata for the audio collection.

The collection-level metadata (Figure 6) consists of the 15 Dublin Core [1] elements (with some extra contributors such as Transcriber, Graphic Artist and Coordinator) plus a Size element to store the size of the collection. The Dublin Core Element Set (with extensions) was chosen for the top level metadata description to enable Dublin Core-based search engines to find the collection's Home Page

The interview-level metadata input form includes fields for the Interviewer, Interviewee and Transcriber contributors as well as the Date and Place at which the inter-

view took place. The recording-level metadata input form provides fields to specify the file format details for each of the digitized audio recordings. It includes: format (mime type), duration, channels, sampling_rate, sample size. Screen dumps of these input forms are illustrated in [32].

BCollection Level Metadata	
Title:	From Lunchroom to Boardroom
Creator:	Therese Collie
Subject:	Women in Trade Unions -- Australia Labor Movement -- Australia -- History
Description:	This project combines oral, written and visual art techniques to highlight the achievements of older women in the labor movement in Queensland between 1930 and 1970.
Publisher:	Trades and Labor Council of Queensland
Contributors	
Transcriber:	Sue Pechy
Graphic Artist:	Judith Hewitson
Coordinator:	Mary Edwards
Date:	1991
Type:	sound.speech.oral_history
Format:	40 x 26 min AGFA Professional Tapes plus 4 x A3 posters
Identifier:	http://sunspot.dstc.edu.au:8888/
Source:	University of Qld Fryer Library UQFL 300
Language:	en
Relation - Has Parts:	Interviews 1-16
Coverage:	Queensland 1930-1970
Rights:	Trades and Labor Council of Queensland
Size:	16 interviews
OK Reset Cancel	

Fig. 6. Metadata Input Form for the Collection Metadata

In addition, the application enables users to open a digital audio (MP3) file and either load the corresponding transcript file (generated manually or using automatic speech recognition techniques) or to enter the transcript interactively. The finer the granularity of temporal indexing, the greater the precision of search and retrieval capabilities but the higher the cost of the indexing procedure.

For this application we aligned each Question-and-Answer in the interview to its corresponding audio segment. Figure 8 shows the transcript window with the audio play-back bar at the top and the currently-selected transcript which is highlighted. Figure 7 shows the Segment metadata window in which the speaker names, keywords and start and end time stamps are specified. Default values for speaker names and keywords can be retrieved from the corresponding Interview metadata. Start and end times are retrieved automatically from the current position of the audio playback bar.

(B)Segment Level Editor

Identifier: Segment 6

Speakers: Therese Collie, Bessie Lymburner

Keywords: Aboriginal Working Conditions, Palm Island

Set Start Time Set End Time

OK Reset Cancel

Start Time: 0:7:30.772
End Time: 0:13:1.355
Duration: 0:5:30.582

Fig. 7. Segment Level Metadata Input Form

(B)lyburn

File Edit View

00:13:01.35 / 00:24:13.32

Close Audio File

TERESA: What was your husband's name?

BESSIE: Eric Lymburner. He was born in Ayr and he had his education down here. He came over to Palm Island when he was 15 year old. You might as well say that he went from freedom to punishment there. He found it hard living, and when he grew up couldn't take it any more. You know, just went and asked, spoke to managers, asked if they can increase the wages and better homes. The strike was on in 1957, and we got pushed off there.

TERESA: So Eric and some other men, five others, were chained and - handcuffed - and sent to-----

BESSIE: Across the sea, over to Townsville, and my husband told me it was pretty rough. They was vomiting on one another, and they didn't take the handcuffs of him until they reached Townsville. Then they had them locked up and the key, they locked them up in the watch-house, never let them out. Then next day they sent me off the island, me and my children. I had three girls and one boy. They sent us off to go to another settlement, and they sent three to Cherbourg and three to Moorabinda, but I didn't stay there long. I was

Fig. 8. Transcript Editor and Alignment Interface

4.2 Saving the Metadata

One of the fundamental assumptions of the ABC model is that no single metadata schema is the correct one and that different users or application contexts require different views of the same set of objects, events and relationships. The underlying data model enables any of these different views to be generated from the underlying description, without lossiness, misinterpretation or ambiguity. For this particular application, there are a large number of possible views or output options for saving the metadata descriptions. Each of these alternatives offers particular advantages within certain contexts:

- ID3 tags [12] embedded within the MP3 files - this approach enables the MP3 files to be found by most generic MP3 Internet Search Engines. It also has the advantages associated with encapsulating the metadata and the actual object together in a single container;
- HTML Pages with metatags [15] in the headers - this approach enables either the collection or individual interviews to be found by generic Web Search Engines such as Yahoo, Altavista or HotMeta [22];
- Dublin Core (DC) [1] descriptions encoded in RDF [13] - this approach enables interoperability with other DC-based systems or users. It has the advantage of being media-independent;
- MPEG-7 Description Schemes [14] encoded in XML [17] - this approach enables interoperability with other media-specific MPEG7-based systems or users;
- Relational tables in a backend database linked to the MP3 files - this approach provides a fast, scalable but application-specific web search interface with high precision and recall.

4.3 Implementation of the Schema Mappings

Currently these mappings are being done programmatically, however we are investigating more generic, flexible and scalable approaches, including:

- merging a XML Infosets into a single composite Infoset [23, 24];
- using Equivalence classes within XML Schema Language to define mappings [19];
- using XSLT (XSL Transformation Language) [25] to transform an XML description from one domain to another.

5 The Search and Browse Interface

A Web front-end [9] was built to enable Internet access to the oral history interviews (transcripts and audio files) and the posters belonging to the *From Lunchroom to Boardroom* project. Through this interface, users can search the collection for particular speakers, words in the transcripts or keywords (the subject of each interview/segment). Figure 9 illustrates the search interface.

Each search request is passed to a Java Servlet which connects to the Oracle database using JDBC and invokes an SQL statement reflecting the search request. When a match is found, the following information is returned:

- The context of the matching word/keyword (either the complete Q&A or the surrounding 200 words, whichever is smaller);
- The interviewee of the segment which contains the match;
- Unique IDs for the recording, interview and collection which contain the match;
- The duration of the matching segment (hrs:mins:secs.millisec);
- URLs to:
 1. the transcript for the segment;
 2. the MP3 audio segment which contains the match;
 3. the MP3 audio file for the complete recording which contains the match.

The search results page (Figure 10) contains links to these URLs via three image buttons. If any of these hyperlinks are selected another Java servlet is invoked which connects to the database and retrieves either the transcript, a segment of the MP3 file or the complete MP3 file (corresponding to a complete recording). If an MP3 audio file or segment is requested, then the currently configured MP3 player plugin is invoked. Recommended MP3 players include: MXAudio (Solaris), Winamp (WinNT), RealPlayerG2 (Solaris, WinNT, Win97).

Fig. 9. Web Search Interface to the MP3 Oral History Archive

6 Comparisons, Conclusions, and Future Work

6.1 Comparisons and Conclusions

Although there are quite a large number of oral history sites on the Internet [26, 27], they generally only provide access to transcripts and only occasionally do they provide sample audio clips in RealAudio, WAV or MP3 formats [28,29]. The authors of this paper believe that this project is the first online demonstration of an oral history archive which enables the fine-grained retrieval and replay of MP3 audio segments based on word or subject searches. Unlike other similar audiovisual retrieval systems [31], this application is based on a metadata schema and model which will enable interoperability between domain-specific standards.



"FROM LUNCHROOM TO BOARDROOM"

Search Results

Your search returned 8 results, which are presented in an unranked order.

Transcripts found by keywords: Aboriginal

...our people used to go out to work, and the same thing happened. They'd just give them a couple of weeks order, and tell them they're broke. I believe the government at that time was the Trustee for Aborigines, even the people that was on the mainland, the men used to work cattle stations and the women did the same work, washing and ironing, cooking. They were treated the same. They were us...

[More text is available - click: Read Transcript]

Collection: 1
Interview: Interview 15
Interviewee: Bessie Lymburner
Recording: Reel 37

choose from the options on the right to view/hear more information. NOTE: You need an installed mp3 player to listen to the audio.

Result 1
Duration (hr:min:sec)
0:2:34.960

[Read Transcript](#)
[Play Segment](#)
[Play Recording](#)

Fig. 10. Results of Keyword Search on 'Aboriginal Health'

The MEGGIE application enables efficient, accurate, economically-viable, fine-grained indexing of audio files through the computer-assisted input of multi-level metadata descriptions and temporal alignment of audio with transcripts and keywords. This indexing approach enables the search and retrieval of very specific and precise audio content based on particular words or keywords in the transcript. This demonstrator will greatly increase the profile, accessibility and dissemination of the valuable historical resources contained in the *From Lunchroom to Boardroom* collection.

The use of the underlying ABC event-aware model enables the recorded metadata to map to any of the five domain-specific vocabularies which are applicable in this context, without ambiguity or loss of integrity or information.

Although the ABC model is still undergoing refinement, this project has shown that the fundamental concept of ABC, which is the explicit recognition of *events* (and their properties) in resource lifecycles, is essential to the generation of interoperable metadata descriptions, particularly in the context of multimedia collections.

6.2 Future Work

The plans for future work include:

- Replacing the XML DTD with an XML Schema;
- Integrating an XML Schema parser into MEGGIE and using the DOM API to browse and edit the structured metadata descriptions;
- Investigating non-procedural mechanisms for mapping between schemas e.g.
 - XML Infosets;
 - XML Schema Equivalence classes;
 - XSLT (XSL Transformation Language);
- Further refinement and extension of the ABC model to support the explicit modeling and recording of intellectual property rights metadata;
- Integrating the MEGGIE, PEGGIE [16] and VEGGIE [30] applications to create a media-independent metadata editor/generator.

Acknowledgements.

The authors wish to acknowledge the use of material belonging to the Trades and Labor Council of Queensland in this project. We also wish to acknowledge the valuable assistance provided by Therese Collie, Niles Elvery (SLQ), Gulcin Cribb and Ros Follett (UQL). The authors also wish to acknowledge the valuable contributions which the Harmony project collaborators, Carl Lagoze and Dan Brickley have made to this work. The authors also wish to acknowledge that this work was carried out within the CRC for Enterprise Distributed Systems Technology established under the Australian Government's Cooperative Research Centre (CRC) Program and acknowledge the support of CITEC and DISR's Technology Diffusion Grant.

References

1. Dublin Core Metadata Initiative, <http://purl.org/DC>
2. International Federation of Library Associations and Institutions, Functional Requirements for Bibliographic Records, March (1998) <http://www.ifla.org/VII/s13/frbr/frbr.pdf>
3. ICOM/CIDOC Documentation Standards Group, CIDOC Conceptual Reference Model,
4. <http://www.ville-ge.ch/musinfo/cidoc/oamodel/>
5. D. Bearman and K. Sochats, "Metadata Requirements for Evidence." Archives & Museum Informatics, University of Pittsburgh, School of Information Science, Pittsburgh, PA (1996) <http://www.lis.pitt.edu/~nhprc/BACartic.html> Overview of the MPEG-7 Standard <http://drogo.cselt.stet.it/mpeg/standards/mpeg-7/mpeg-7.htm>
6. INDECS Home Page <http://www.indecs.org/>.
7. D. Bearman, G. Rust, S. Weibel, E. Miller, J. Trant, A Common Model to Support Interoperable Metadata. D-Lib Magazine, January (1999) <http://www.dlib.org/dlib/january99/bearman/01bearman.html>
8. The Harmony Project, <http://www.ilrt.bris.ac.uk/discovery/harmony/>

9. C. Lagoze, J. Hunter, and D. Brickley, An Event-Aware Model for Metadata Interoperability, ECDL2000, Lisbon, September 2000
10. MPEG.ORG, The MP3 Portal, <http://www.mpeg.org/MPEG/mp3.html>
11. From Lunchroom to Boardroom: Stories and Images of Women's Achievements in the Labor Movement 1930's - 1960's, <http://sunspot.dstc.edu.au:8888/>
12. ID3v2 Home Page, <http://www.id3.org/>
13. O. Lassila, R. Swick, Resource Description Framework: (RDF) Model and Syntax Specification, W3C Proposed Recommendation PR-rdf-syntax-19990105, January (1999) <http://www.w3.org/TR/PR-rdf-syntax/>
14. MPEG-7 Multimedia Description Schemes Working Draft Version 2.0, ISO/IEC JTC1/SC29/WG11/N3247, Noordwijkerhout, March (2000)
15. John Kunze, Encoding Dublin Core Metadata in HTML, Informational Internet Draft, May (1999) <http://www.ietf.org/internet-drafts/draft-kunze-dchtml-01.txt>
16. J. Hunter, Z. Zhan, An Indexing and Querying System for Online Images Based on the PNG Format and Embedded Metadata, ARLIS/ANZ Conference, Brisbane, Australia, September (1999) <http://archive.dstc.edu.au/RDU/staff/jane-hunter/PNG/paper.html>
17. Extensible Markup Language (XML) 1.0, REC-xml-19980210, W3C Recommendation 10 February (1998) <http://www.w3.org/TR/REC-xml>
18. D. Brickley, R. V. Guha, Resource Description Framework (RDF) Schema Specification, W3C Candidate Recommendation CR-rdf-schema 20000327, March (2000) <http://www.w3.org/TR/2000/CR-rdf-schema-20000327/>
19. H. S. Thompson, D. Beech, M. Maloney, N. Mendelsohn, XML Schema Part 1: Structures, W3C Working Draft April (2000), <http://www.w3.org/TR/xmlschema-1/>,
20. P. V. Biron, A. Malhotra, XML Schema Part 2: Datatypes, W3C Working Draft WD-xmlschema-2-2000025, <http://www.w3.org/TR/xmlschema-2/>, April 7 (2000)
21. Document Object Model (DOM), <http://www.w3.org/DOM/>
22. HotMeta Metadata Search Engine, <http://www.dstc.edu.au/Research/Projects/hotmeta/>
23. J. Cowan, D. Megginson, XML Information Set, W3C Working Draft WD-xml-infoset-19991220, <http://www.w3.org/TR/xml-infoset>, December 20 (1999)
24. J. Marsh, D. Orchard, XML Inclusions, W3C Working Draft WD-xinclude-20000322, <http://www.w3.org/TR/2000/WD-xinclude-20000322>, March 22 (2000)
25. J. Clark, XSL Transformations (XSLT), W3C Recommendation REC-xslt-19991116, <http://www.w3.org/TR/xslt>, November 16 (1999).
26. The Center for Oral History at the University of Connecticut Links to Oral History Sites, <http://www.ucc.uconn.edu/~cohadm01/links.html>
27. The Department of History, University of California, Riverside, Links to Oral History Resources, <http://www.ucr.edu/h-gig/hist-topics/oralh.html>
28. Many Voices One Story: Listening Room, BGC Archives, Wheaton College, <http://www.wheaton.edu/bgc/archives/exhibits/ohistory/oral03.htm>
29. American Memory Sound Recordings, Library of Congress, <http://memory.loc.gov/ammem/audio.html>
30. J. Hunter, J. Newmarch, An Indexing, Browsing, Search and Retrieval System for Audiovisual Libraries, ECDL '99, 22-24 September, Paris. <http://archive.dstc.edu.au/RDU/staff/jane-hunter/ECDL3/paper.html>
31. B. Simpson-Young, K. Yap, Trialing a System for Remote Navigation of Film Archives, SPIE International Symposium on Voice, Video and Data Communications, Boston, MA, November (1996)
32. J. Hunter, D. James, Application of an Event-Aware Metadata Model to an Online Oral History Project, <http://archive.dstc.edu.au/rdu/staff/jane-hunter/OralHistory/paper.html>

From the Visual Book to the WEB Book: The Importance of Good Design

M. Landoni¹, R. Wilson², and F. Gibb¹

¹Department of Information Science, University of Strathclyde.

²CDLR: Centre for Digital Library Research, University of Strathclyde.

Abstract. This paper presents the results of two studies into electronic book production. The Visual Book study [1] explored the importance of the visual component of the book metaphor for the production of more effective electronic books, while the WEB Book study [2] took the findings of the Visual Book and applied them to the production of books for publication on the WWW. Both studies started from an assessment of which kinds of paper book are more suitable for conversion into electronic form, and both identified as target publications those which are meant to be used for reference rather than those which are read sequentially and usually in their entirety by users. This includes scientific publications and textbooks which have been chosen for the Visual Book and the WEB Book experiments. In this paper we discuss the results of the two studies and the way they could influence the design and production of more effective electronic books.

1 Introduction

The result of integrating the classical book structure with features which can be provided within an electronic environment constitutes a generally accepted definition of an electronic book. The main features of electronic books are that they are dynamic, reactive, can be made available in different formats and/or editions in a short time and, with the growth of the Internet, are accessible almost everywhere [3],[4]. However the cognitive overhead associated with a computer-based environment is an important reason for carefully considering the appropriateness and the method of realising this conversion [5],[6]. It is important to consider both the subject matter and the intended usage of a particular paper book in order to decide whether an electronic version will be useful or not [7], [8]. It seems clear that the desire or willingness to use electronic books is primarily based on the need for rapid access to information chunks rather than for assimilating significant amounts of narrative text [9]. An analysis of the market for paper and electronic publications showed that users prefer to keep a paper version in addition to the electronic one even when a free electronic version is available on WWW [10].

The motivations for producing electronic books therefore deal with factors related to the market requirement for electronic books, as well as cognitive issues related to the ability of the reader to use, appreciate and prefer books in electronic format to paper ones [11]. The quality aspect becomes crucial and for this reason the design process has to be reviewed in order to create effective electronic books, i.e. communicators of

electronic information which have additional value which paper cannot provide. This paper describes two studies into the design and production of electronic books: the Visual Book and the WEB Book. The next three sections discuss the theory behind the Visual Book experiment, its design and findings: this provides the background for the description of the WEB Book experiment. Finally conclusions are drawn from both experiences.

2 Visual Rhetoric

The Visual Book experiment was a study into the application of the book metaphor to the design and production of electronic books. The novel aspects of the Visual Book studies was the importance given to the visual components of the physical book when designing electronic book. The idea behind visual rhetoric [14] is to use visual clues to identify or highlight those parts of a document which are more important for the comprehension of its meaning. Previous research has corroborated this view that a document can be interpreted as a visible representation of a text according to its semantic contents [15]. Visual rhetoric provides the reader with a graphical mark up language which is immediately recognisable on the basis of previous reading activity. Different graphical presentations suggest different readings and deeply affect the interpretation of the contents of the same text. The Visual Book project focused on the importance and use of visual rhetoric when presenting information on a screen and in particular on the influence of visual rhetoric in the presentation and use of electronic books.

3 The Visual Book Experiment

The Visual Book represents a particular interpretation of the electronic book, based mainly on the visual aspects of the paper book such as dimensions, thickness, page form and general design style [1]. A Visual Book is the result of the process of converting an existing paper book into electronic form using two main components:

- The Visual Book Builder: the authoring system for building the Visual Book.
- The Visual Book Browser: the module for presenting the Visual Book to readers.

The physical aspect of the paper book is the key to the construction and representation of a Visual Book. Emphasis is placed on preserving the book's appearance in order to draw upon the familiarity the reader already has with books in general. Book functionalities were studied in terms of how books are used in practice to ensure that information is presented in a natural and familiar way and that the reading process is supported in an amenable environment.

The main idea of a Visual Book is to offer the reader an object as similar as possible to the paper book by replacing its essential physical features, such as size and quality, with visual ones [16]. Particular attention was paid to identifying those functionalities a Visual Book could acquire from electronic support which would be most useful and consistent with the specification of the book concept. One of the main advantages of electronic support is the ease with which the designer can modify, update and distribute data [17]. On the other hand there is a series of functionalities, which

already form part of the paper book environment, that are supported only in part by other electronic book models. These include:

- Bookmarks.
- Notes on the margins or elsewhere in the text.
- Highlighting of interesting parts.
- Easy access to pages which are frequently consulted by following different visual clues.
- Information about the ratio between the pages already read and those remaining.
- Control of the reading progress.
- Browsing and scanning for interesting sections.

All of these have been incorporated as essential features of the Visual Book Browser, as shown in Figure 1. In particular search tools, such as an Index tool and Table of Contents (ToC), have been added to the Visual Book features.

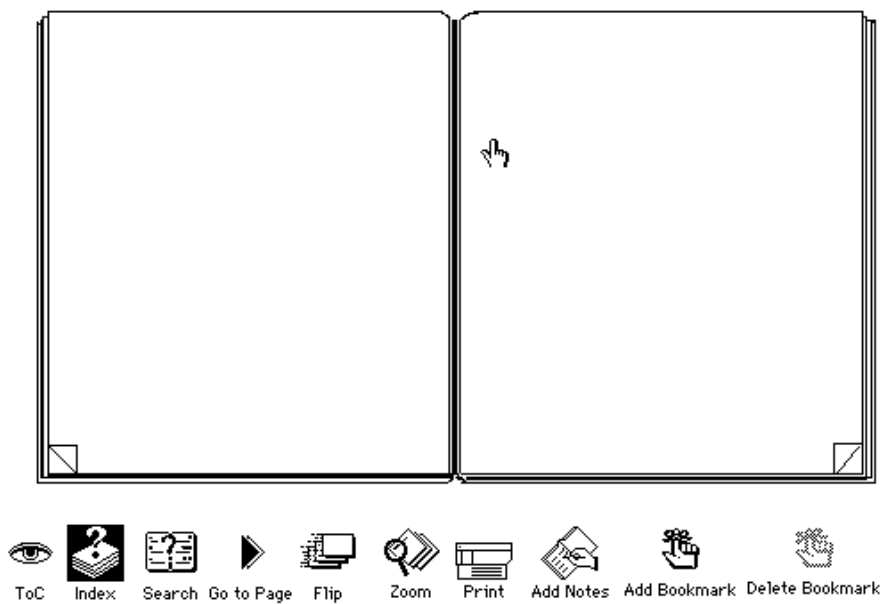


Fig. 1. The Visual Book Browser

4 Visual Book Findings and Analysis

The evaluation of the Visual Book [1] concentrated on aspects of the book metaphor when converting paper books into electronic format. In particular the importance of maintaining the same typographical features, the design rules, the pagination format and the paper appearance were considered carefully in terms of retaining the meaning of the original text. The physical aspects of the page are encapsulated in a set of presentation rules through which visual clues give the reader semantic information about the context. This primary feature of the book metaphor has been considered to

be particularly important and to be an original contribution to the future development of the electronic design process.

A number of criteria were extracted from other evaluations of the processes used when reading from a computer [18], [19] and these were used as the basis for the Visual Book evaluation:

- **Sense of directness** is the degree of feeling users have that changes on the screen are the result of their actions. It is connected with the illusion users have that the displayed image is a physical object which can be manipulated in the same way as a real paper book.
- **Sense of engagement** is the level of interest the system induces in users. One source of engagement is the fun of seeing the system react and is related to the novelty of the system; tangibility and responsiveness are also responsible for a good level of engagement. Paper has generally a low level of engagement because it is not interactive and is overly familiar to users.
- **Sense of text** is the feeling users may have of the structural and semantic structure of the text that is being read, i.e. its spatial disposition. Readers are known to be able to recall the position of text in a paper text [20]. Factors which can influence the sense of text are the page size, legibility, and a low responsiveness while scrolling when looking for more text.

In general the results of the evaluation of the Visual Book, which were supported by the findings of a similar project - the Hyper-Book [21], showed that the book metaphor was both accepted and understood by its evaluators. These results also highlighted the necessity for a new role in electronic publishing, "the designer of electronic books, as the person in charge of the final appearance of the electronic book". This is a person competent not only in the technological aspects of producing electronic documentation but with a background in understanding the importance of presentation issues such as pagination and the general format and appearance of the electronic document, which can be called collectively electronic typography. This means that typographical rules are essential components of the cognitive model of a book and hence relevant to the conversion of the metaphor of the book into electronic form.

The evaluation method which was applied to the Visual Book was the cognitive jog-through [22]. In this approach, users were free to express, in written or verbal form, their comments and suggestions. Users were happy to interact with an object which resembled a book and appreciated that its enhanced functionalities were consistent with the original paper version. The main request was for an intelligent search function to combine the precision provided by paper visual clues and by index terms with the high recall provided by full text search which can be performed very efficiently using mainstream information retrieval software.

In addition to the comments from evaluators a number of future developments were identified by the research team based on their experiences. In particular the influence of the Internet on the presentation of information in electronic format should be studied in order to understand if modifications to the book metaphor and the related concept of visual rhetoric have to be considered to make it compatible with the Internet paradigm. It is this issue that the WEB Book investigated and which is described in the next section.

5 The WEB Book Experiment

The WEB Book was based on the central hypothesis of the Visual Book project: that appearance is an important factor in the effective presentation of information on screen, and applied it to the production of books on the WWW. The aim was to study whether focusing on the appearance of the content when preparing a scientific textbook for electronic publication has a positive impact on its usability.

A series of studies by Morkes and Nielsen [12],[13] into how to write for the Web provided the basis for a methodology which would enable any improvement to be measured. They proposed that the usability of a text on the Web can be greatly increased by altering the presentation of the information to make it concise, scannable and objective. Their definition of what concise, scannable and objective text is as follows:

- **Concise** text is where less-important information is edited out to reduce the page length.
- **Scannable** text is written to encourage scanning, or skimming, of the text for information of interest. For this purpose bulleted lists, boldface text to highlight keywords, photo captions, shorter sections of text, and more headings, etc. are used.
- **Objective** text presents information without exaggeration, or making subjective or unsustainable claims.

When applied in their own experiments to a site with tourist information and a site with technical white papers (i.e. papers in progress), these changes led to increases in usability of 124% and 159% respectively. The WEB Book project adhered to their procedure, as described in *Applying Writing Guidelines for Web Pages* [13] as closely as possible, and replicated the content of the questionnaires from their third study, using Chapter Five (Search Strategies) of van Rijsbergen's *Information Retrieval* [23] as the material for the experiment¹. This textbook is one of the fundamental readings of IR, and a relatively plain machine readable version of the book already exists on the Internet.

Some modifications to Morkes and Nielsen's original procedures were necessary to suit the purposes of this research². Because a scientific textbook is primarily objective in nature, and cannot be made concise without either breaching copyright or compromising the meaning of the text, the emphasis in the WEB Book was simply on reworking the chapter to make it scannable by improving/redesigning its logical and physical layout. As the experiment was being conducted on a different type of material, its objectives varied from those of the original studies, so that browsing rather than searching for information in the text was encouraged, making measurement of the time taken to complete the tasks less central to the purpose of the study.

The resulting methodology was used to test three hypotheses:

- **Hypothesis 1:** Users of the revised, scannable version of the chapter will make fewer errors on tasks than will users of the original version.

¹ Permission to use this material by the author is gratefully acknowledged.

² Revised version of the questionnaires used can be found at:
<http://www.dis.strath.ac.uk/students/ruth/forma.html>.

- **Hypothesis 2:** Users of the revised version will report higher subjective satisfaction with the site than will users of the original.
- **Hypothesis 3:** When measures from the first two hypotheses are combined into an overall usability score for each version of the site, the scannable version will have a higher usability score than the control version (i. e. the original electronic version).

Each stage of this experiment was carried out over the Internet. Participants consisted of 18 respondents to emails sent to two mailing lists for the IR community and to a group of postgraduate students who had studied IR in the Department of Information Science, University of Strathclyde. By focusing on the IR community the intention was to target users familiar with the general subject matter of the textbook they would be asked to read, so that they could concentrate on browsing the site for information rather than being distracted by entirely alien content. Although mailing list respondents inevitably comprised a greater number of experts, similar numbers in each group had consulted van Rijsbergen's book before and the students were considered to have sufficient background in the field to be valid subjects in the experiment.

Two electronic versions of Chapter Five were used in this study: the original ³, and a revised version, written for the purpose of the experiment ⁴. Of the eighteen participants in the experiment eight looked at the original version of Information Retrieval, and ten looked at the revised version.

6 Design Issues for the WEB Book

In the original electronic version only minor concessions had been made to the new medium. Each chapter is allocated a separate "page" (i.e. a web page) and text is organised in a linear way so that users have to keep scrolling down until they reach the end of the chapter. Hypertext functions provided by Web browsers are used to provide certain basic structural links, such as moving from one chapter to the next. There are links from citations to bibliographic references, and links from the table of contents to the beginning of chapters have been inserted. Otherwise, it does not provide any more effective functionalities than those provided in the paper book.

The revised version of this chapter was prepared for this study by the research team and followed Morkes and Nielsen's guidelines for improved scannability. The chapter was divided into ten separate files (*Introduction*, *Boolean search*, and so on) according to the sub-headings in the original text, so that each sub-section occupied a separate page. In addition, a chapter contents page was created to list and provide links to each section. Hyperlinks were exploited in order to aid navigation. At the top of each page, links were inserted to enable the user to jump back to the previous section, forward to the next section, or back to the table of contents for the chapter. At the foot of each page there was a list containing hyperlinks to each section in the chapter. Extra headings, coloured and sized, were inserted in the text in order to divide it into smaller, more digestible chunks. All diagrams were centred to give the pages a

³ Available at: URL <http://www.dcs.gla.ac.uk/Keith/Chapter.5/Ch.5.html>.

⁴ Available at: URL <http://www.dis.strath.ac.uk/students/ruth/chpt5/>.

neater, more symmetrical appearance, and in order that they could be easily distinguished from the text. Key words and phrases were coloured red to call attention to important pieces of text, and lists were indented, their numbers or bullets also coloured red. These changes were intended to make the chapter more easily scannable by giving it a design which exploited more fully the functions of the Web as a medium.

7 The WEB Book: Findings and Analysis

The two major measures used to calculate the overall usability of each version of the chapter were the percentage of errors made in the search tasks, and subjective satisfaction (the culmination of all the individual elements of the questionnaire). These results are presented in Table 1.

Table 1. Mean scores for the two major measures, with standard deviations in parentheses. (Task errors are percentage scores, and subjective satisfaction are the mean scores from the questionnaires, on a scale from one to 10).

Condition	Task Errors	Subjective Satisfaction
Original	16.7 (17.8)	5.1 (1.4)
Revised	6.7 (14.0)	7.6 (0.8)

To determine how much better or worse in percentage terms the revised chapter was relative to the control, Morkes and Nielsen's procedure for normalising the participant groups' mean scores for the two measures was adopted. For each measure, the revised condition's mean scores were transformed by division relative to the control (see Table 2) so that scores above 100 are better than the control, and those below 100 are worse.

Again following the original experiment, the overall usability score for each version was calculated by taking the geometric mean of the normalised scores for the two measures. (The geometric, rather than arithmetic, mean was used because ratios were being compared). Table 2 shows that 150% fewer task errors were made by users of the revised version, and these participants recorded 48% higher subjective satisfaction. Combined, this means an increase in overall usability of 92%. The chapter which was rewritten to take into account the appearance of the content so that it was more scannable, achieved a higher score in every single measure than the original version of the text, and the three hypotheses were substantiated:

- **Hypothesis 1** was verified. Users of the revised version made 150% fewer errors on tasks than users of the original version.
- **Hypothesis 2** was supported. Users of the revised version reported 48% higher subjective satisfaction than users of the original version.
- **Hypothesis 3** was confirmed. Overall usability scores for the two versions of the site show that, compared to the control version, the revised version was 92% better.

Table 2. Normalised mean scores for the two major measures and overall usability.

Condition	Task Success ⁵	Subjective Satisfaction	Overall Usability
Original	100	100	100
Revised	250	148	192

Participants' comments were in accordance with these results. Users of the original version of the chapter disliked the design of the pages, describing it variously as "flat and not very appealing", "very user-unfriendly" and "plain", with formulae that are difficult to read, making it hard to concentrate on searching for information. Users of the revised version, on the other hand, made some positive remarks about the page design, describing it as "easy to navigate" and "clean-looking". Problems with the site were also identified, with one participant commenting on the problem of having to scroll up and down to compare the different diagrams and equations.

In this case study, making the chapter more scannable according to Morkes and Nielsen's guidelines led to a 92% increase in usability, indicating the positive impact of focusing on the appearance of the content when preparing a scientific textbook for publication on the Web. Although the experiment was conducted on a small scale, these results are indications of the potential effect on usability of altering the appearance of texts on the Web to make them more scannable. It would benefit from being replicated on a larger scale, perhaps on the whole textbook, and with users from a wider range of backgrounds. Nonetheless, these indications are positive and in tune with the findings of previous exercises which form the background to the experiment. They reinforce the hypothesis of Crestani and Melucci's Case Study of Automatic Authoring [24] in which it is suggested that textbooks, because they are often used for accessing specific facts and pieces of information, are more effectively represented electronically in a non-linear fashion. Further, this research is in accordance with a key premise of the Visual Book project, which states that a book's appearance is a main factor in the specification of interfaces designed to present information which has been published on paper.

8 Conclusions

The aim of the two projects described in this paper is to promote the definition of guidelines for the production of more effective electronic books. The definition of what an electronic book is, or has to be, is quite loose and unfortunately this is unavoidable because of the nature of the object and the difficulty of defining something with so many aspects and functionalities. In general the perspective of the whole Visual Book and WEB Book projects is centred on readers and their needs, expectations, cognitive models and requirements. The results of both experiments have shown that the book metaphor plays a crucial role in the design of electronic books.

⁵ This was "Task Error" in Morkes and Nielsen original study, but it was changed to avoid misunderstanding.

In particular:

- The page metaphor should be respected. That is, the page should be treated as a visual space where information can be easily found and scanned.
- The logical structure of the book has to be considered. Table of Contents and Index(es) are essential features. These cannot be simply replaced by search facilities whose complexity makes the reading process more confusing for the user.
- The book template has to be used strictly to present information which is book related, and not any other kind of material as the result would be a heterogeneous system which could confuse users.
- Titles, pagination, typographical aspects have to be designed carefully to enhance text readability and scannability.
- Visual clues have to be adapted to exploit the potential of the medium where the book is published.

These observations were originally derived from the Visual Book experience but have proven to be flexible enough to be applied to the WEB Book. In fact they could be summarised in what Morkes and Nielsen [12],[13] recognised as the need for scannable text when the electronic book is going to be delivered on the WWW. Presentation and appearance play a crucial role in allowing users to read and use text in electronic form as much as they did and do with information on paper. Thus it is not surprising that guidelines for the production of good electronic books will have to take into account issues related to style and layout. The authors of this paper believe that greater emphasis on the observation of readers' reactions, can lead to the definition of heuristic rules for the production of "good" and "readable" electronic books.

References

1. Landoni, M.: The Visual Book system: A study of the use of visual rhetoric in the design of electronic books. Glasgow: Department of Information Science of the University of Strathclyde (PhD Thesis). (1997).
2. Wilson, R.: The importance of appearance in the design of WEB books. Glasgow: Department of Information Science of the University of Strathclyde (MSc Diss.) (1999).
3. Barker, P.: Interactive Electronic Books. *Interactive Multimedia*, 2(1) (1991) 11-28.
4. Barker, P.: Living books and dynamic electronic libraries. *The Electronic Library*, 14(6) (1996) 491-501.
5. Barker, P., Richardson, S. and Benest, I.D.: Human-computer interface design for electronic books. In: Raitt, D.I. and Jeapes, B. (eds.). *Proceedings of the Online Information 94*, 16th International Online Information Meeting, London 6-8 December. Oxford: Learned Information (1994) 213-292.
6. Benest, I.D., Morgan G., and Smithurst M.D.: A humanised interface to an electronic library. In: Bullinger, H. J. and Shackel, B. (eds.): *Proceedings of INTERACT'87*. Amsterdam: Elsevier Science (1987) 905-910.
7. Wilson, T.: Electronic publishing and the future of the book. Presented at the International Conference on Book Science, Vilnius University, Lithuania. (1997). Available at: <http://www.shef.ac.uk/~is/publications/infres/paper39.html> (Last visited 1/7/99).
8. Landoni M., Crestani F. and Melucci M.: The Visual Book and the Hyper-TextBook: Two Electronic Books One Lesson?. In: *Proceedings of RIAO'2000 Conference on Content-based Multimedia Information Access*, Paris, April 12-14. Paris: C.I.D. (2000) 247-265.

9. Leaver, R.: The future of electronic publishing for book publishers within Britain. *Aslib proceedings*. 47 (7/8). July/August. (1995) 164.
10. Pope, B. K.: How to succeed in online market- National Academy Press: A Case Study. *The Journal of Electronic Publishing*. June. 4(4) (1999). Available at: URL <http://www.press.umich.edu/jep/04-04/pope.html>.
11. Catenazzi, N., Landoni, M. and Gibb, F.: Design issues in the production of hyper-books and visual-books. *Alt-J Association for Learning Technology Journal* 1(2) (1993) 40-54.
12. Morkes, J. and Nielsen, J.: Concise, SCANNABLE, and objective: how to write for the Web. Available: URL <http://www.useit.com/papers/webwriting/writing.html> (Last visited 11/6/99) (1997).
13. Morkes, J. and Nielsen, J.: Applying writing guidelines to Web pages. Available: URL <http://www.useit.com/papers/webwriting/rewriting.html> (Last visited 11/6/99) (1998).
14. Landoni M. and Gibb F.: The Importance of Visual Rhetoric in the Design and Production of Electronic Books: The Visual Book Experience. *The Electronic Library*. In press. (2000).
15. Southall, R.: Interfaces between the designer and the document. In: André, J., Furuta, R. and Quint, V. (eds.). *Structured documents*. Cambridge: Cambridge University Press (1989) 119-131.
16. Landoni, M., Catenazzi, N. and Gibb, F.: Hyper-books and visual-books in an electronic library. *The Electronic Library*, 11(3) (1993) 175-186.
17. Catenazzi, N. and Gibb F.: The publishing process: the hyper-book approach. *Journal on Information Science*. 21(3) (1995) 161-172.
18. Hansen, W.J. and Haas, C.: Reading and writing with computers: a framework for explaining differences in performance. *Communication of ACM*, 37(9) (1988) 1080-1089.
19. Egan, D.E., Lesk, M. E., Ketchum, R. D., Lochbaum, C. C., Remde, J. R., Littman, M., and Landauer T. K.: Hypertext for the Electronic Library? CORE Sample Results. In: *Proceedings of Hypertext '91*, San Antonio. New York: ACM Press (1991) 299-312.
20. Rothkopf, E. Z.: Incidental Memory for location of information in text. *Journal of Verbal Learning and Verbal Behaviour*, 10(6), (1971) 608-613.
21. Catenazzi, N.: A Study into electronic book design and production: Hyper-Book and Hyper-Book Builder. Glasgow: Department of Information Science of the University of Strathclyde (PhD Thesis) (1994).
22. Rowley, D. E. and Rhoades, D. G.: The cognitive jog-through: a fast-paced user interface evaluation procedure. In: *Proceedings of CHI'92 Conference on Human Factors in Computing Systems*, Monterey, May 3-7. New York: ACM Press (1992) 389-395.
23. van Rijsbergen, C.J.: *Information retrieval*. 2nd Edition, London: Butterworths (1979). Available at: URL <http://www.dcs.gla.ac.uk/Keith>. (Last visited 30/9/99).
24. Crestani, F. and Melucci, M.: A case study of automatic authoring: from a textbook to a hyper-textbook. *Data and Knowledge Engineering*. 27(1) (1998) 1-30.

Topic Detection in Read Documents

Rui Amaral and Isabel Trancoso

IST, Instituto Superior Técnico
INESC- Instituto de Engenharia de Sistemas e Computadores
Rua Alves Redol, 9, 1000 - LISBOA
email: ramaral@speech.inesc.pt,
Isabel.Trancoso@inesc.pt

Abstract This paper addresses the problem of topic annotation in the speech retrieval domain. It describes an algorithm developed to perform automatic topic annotation of broadcast news (BN) speech corpora. The adopted approach is based in Hidden Markov Models (HMM) and topic language models, solving the topic segmentation and labelling tasks simultaneously. To overcome the lack of topic labelled material for training statistical models, a two-stage unsupervised clustering was developed. Both stages are based on the nearest-neighbour search method, using the *Kullback-Leibler* distance. On-going experiments to evaluate the system performance are also described.

1 Introduction

In the past, most of the research done in the area of topic detection focused on textual media, obviously more prominent at the time. Nowadays, the technology allows us to extend the data storage to others media such as video, images, audio and speech. If we take into account that a great amount of information is spoken (e.g. radio and TV), it is easy to understand the motivation for the recent interest in this area by the speech research community.

The speech retrieval problem cannot be equated in the same way as text retrieval because it is not yet possible to produce perfect transcriptions of speech messages automatically. For example, a state-of-the-art continuous speech recognition system for broadcast news generates automatic transcriptions with a word error rate (WER) of 30-35% [1]. In speech retrieval, the problem is to organise the recordings into a set of categories or classes predefined and known *a priori*, in order to reduce the number of documents that need to be searched in information retrieval systems. The most successful approaches, nowadays, use diverse classification techniques to process the transcriptions generated automatically by Large Vocabulary Continuous Speech Recognition systems (LVCSR)[2].

The next section describes a system developed to perform topic annotation of automatic broadcast news transcriptions (radio, TV). A preliminary test of this system was done with a spoken corpus of read newspaper items. The test and the corresponding results are described in Section 3. Section 4 concludes with our perspectives for future research.

2 System Outline

Figure 1 shows a block diagram of the topic annotation system. The training process depends on the availability of topic-labelled material. If topic-labelled data is available, the topic language models can be created directly from the available transcriptions. Otherwise, the unlabelled corpus is first automatically clustered into topics, and topic language models are constructed for each of the resulting clusters.

The main purpose of the clustering procedure is to join stories in clusters (topics) according to their word similarities. The clustering technique we have adopted, similar to the one described in [3], is a two-stage unsupervised clustering based on nearest neighbour search and the *Kullback-Leibler* distance measure. For each cluster, a topic language model is then built from the corresponding stories, using unigram statistics, which are smoothed versions of a global unigram model obtained with all the training text¹. In the HMM topology we have adopted, each topic is modelled by a unique state with a "self-loop" and a transition probability. The estimation of these probabilities is obtained from the text corpus. The last phase corresponds to the search of the best hypothesis, performed with the *Viterbi* algorithm.

3 Topic Detection Experiments with BD_PÚBLICO

Although planned for the near future, there is still no broadcast news speech corpus for European Portuguese. To overcome this problem, we have used the BD-PÚBLICO corpus [5]. The corpus contains texts from the daily Portuguese newspaper "O PÚBLICO" collected almost on a daily basis from 1995 to 1999, a subset of which was selected for being recorded by 100 speakers.

We have done two types of pilot experiments with this corpus: a first experiment with the textual corpus, and a second one with the much smaller recorded corpus. For the first experiment, a subset of the newspaper texts corresponding to the 28-month period from September 95 to December 97 was used to develop, train, and evaluate the system. This subset is topic-labelled using 9 broad categories: *education (B)*, *culture (C)*, *sports (D)*, *economy (E)*, *science (H)*, *international issues (I)*, *politics (P)*, *media (R)* and *society (S)*. The selection of the stories was performed according to their length, with a maximum of 2000 words and a minimum of 100. After selection, all the texts were pre-processed in order to solve cases such as punctuation and tags.

The first 16 months of the corpus were used to train the system. This training material has about 23 million words spread across 42000 stories and a 560-word per story average length. The remainder of the corpus was equally divided into development and evaluation sets, each containing about 3 million words. The topic annotation results on the evaluation corpus yielded 89.9% correctness and average 260 out-of-vocabulary words (OOV). The topic confusion matrix is shown in Table 1. It shows us that the *media* topic is the most confusable one. It is often confused with the *culture* and *society* topics, which is reasonable since they are strongly related.

¹ The language models were created using the CMU Cambridge Statistical Language Modelling Toolkit [4].

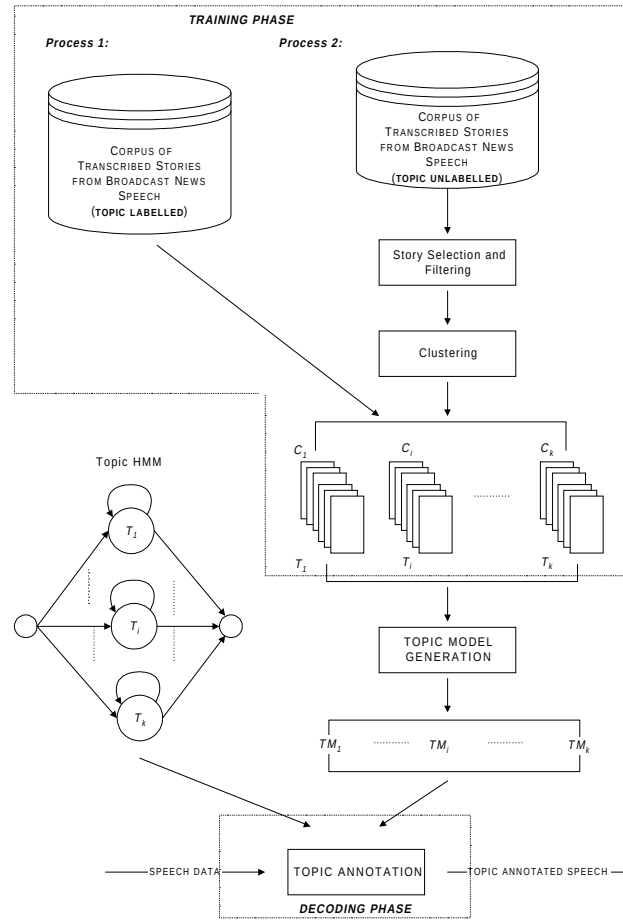


Fig. 1. Architecture of the automatic topic annotation system, for speech messages

Table 1. Topic confusion matrix

	B	C	D	E	H	I	P	R	S	Total	Diag
B	88.7%	3.2%	0.0%	0.0%	0.0%	0.2%	2.1%	0.0%	5.8%	433	89%
C	0.4%	95.7%	0.0%	0.6%	0.0%	0.9%	0.2%	0.5%	1.7%	846	96%
D	0.3%	0.2%	95.2%	0.4%	0.0%	0.0%	1.6%	0.6%	1.9%	1985	95%
E	0.4%	0.3%	0.0%	88.2%	0.0%	0.7%	3.3%	0.0%	7.0%	987	88%
H	1.9%	1.9%	0.0%	1.5%	87.3%	0.0%	0.4%	0.0%	6.9%	259	87%
I	0.0%	0.9%	0.1%	0.6%	0.0%	93.5%	1.8%	0.3%	2.8%	679	94%
P	0.2%	0.7%	0.0%	2.2%	0.0%	3.8%	88.2%	0.0%	5.0%	583	88%
R	1.4%	17.0%	1.7%	4.0%	1.1%	1.7%	2.6%	59.5%	10.9%	348	59%
S	0.4%	2.9%	0.1%	1.7%	3.5%	2.7%	2.1%	0.5%	86.0%	801	86%

The second experiment was done with a subset of the spoken corpus. To do this, the corresponding transcriptions were automatically generated by a Portuguese LVCSR system, with a vocabulary of 5k words and a word error rate of 16.4% [6]. This relatively small subset has only 36 stories, which is the average size of a television broadcast news program. However, they are all very short, corresponding to partial versions of the original stories, with less than 100 words each.

On this subset, the topic annotation system achieved only 61% of correctness. The low score obtained is mainly due to the small size of the lexicon, the consequently large OOV rate and also the small size of the stories.

Conclusions and Future Work

This paper described our first experiments in the area of topic detection for spoken documents. On-going work is mainly focussed on adjusting the referred lexicon and generating a significant number of automatic transcriptions for the full story versions. Our long-term goal, however, is to work on topic detection for broadcast news. This will be done in the scope of the recently started European project ALERT (Alert System for Selective Dissemination of Multimedia Information). For that purpose, we have already started the collection of a pilot corpus of news related programs in cooperation with RTP, the national TV broadcaster, amounting for the time being to approximately 7 hours of video and audio data.

Acknowledgements

This work was done in the scope of project REC - Speech Recognition Applied to Telecommunications, sponsored by FCT. The authors would like to thank their colleagues Diamantino Caseiro and João Paulo Neto for many helpful contributions.

References

1. Fiscus, J., Doddington, G., Garofolo, J., Martin, A., "NIST'S 1998 Topic Detection and Tracking Evaluation (TDT2)", in Proc. DARPA Broadcast News Workshop, Feb. 1999.
2. Ng, K., "Survey of Approaches to Information Retrieval of Speech Messages" Technical report, Spoken Language Systems Group, MIT, February 1996.
3. Yamron, J. P., Carp, I., Gillick, L., Lowe, S., "A Hidden Markov Model Approach to Text Segmentation and Event Tracking", in Proceedings of ICASSP-98, Seattle, May 1998.
4. Clarkson, P., Rosenfeld, R., "Statistical Language Modeling using the CMU-Cambridge Toolkit", in Proc. EUROSPEECH 97, Rhodes, Greece, 1997.
5. Neto, J., Martins, C., Meinedo, H., and Almeida, L., "The Design of a Large Vocabulary Speech Corpus for Portuguese", in Proc. EUROSPEECH 97, Rhodes, Greece, 1997.
6. Caseiro, D., Trancoso, I., "A Decoder for Finite-State Structured Search Spaces" submitted to ISCA ITRW International Workshop on "Automatic Speech Recognition: Challenges for the Next Millennium", Paris, France, September 18-20, 2000.

Map Segmentation by Colour Cube Genetic K-Mean Clustering

Vitorino Ramos and Fernando Muge

CVRM - IST Geo-Systems Centre, Instituto Superior Técnico,
Avenida Rovisco Pais, 1049-001, Lisboa, Portugal
{vitorino.ramos,muge}@alfa.ist.utl.pt
<http://alfa.ist.utl.pt/~cvrm>

Abstract. In this work, a method is described for evolving adaptive procedures for colour image segmentation. We formulate the segmentation problem as an optimisation problem and adopt evolutionary strategy of Genetic Algorithms (GA) for the clustering of small regions in colour feature space. The present approach uses k-Means unsupervised clustering methods into GA, namely for guiding this last Evolutionary Algorithm in his search for finding the optimal or sub-optimal data partition, task that as we know, requires a non-trivial search because of its intrinsic NP-complete nature. To solve this task, the appropriate genetic coding is also discussed, since this is a key aspect in the implementation. Our purpose is to demonstrate the efficiency of GA to automatic and unsupervised texture segmentation. Some examples in Colour Maps are presented and overall results discussed

1 Introduction

Image segmentation is a low-level image processing task that aims at partitioning an image into homogeneous regions [11]. Recently, researchers have investigated the application of genetic algorithms (GA, [13,8,15]) into this problem. Perhaps the most extensive and detailed work on GAs within image segmentation is that of *Bhanu* and *Lee* [3]. Many general pattern recognition applications of this particular paradigm can also be found in [16]. One reason (among others) for using this kind of approach is mainly related with the GA ability to deal with large, complex search spaces in situations where only minimum knowledge is available about the objective function. For example, most existing image segmentation algorithms have many parameters that need to be adjusted. The corresponding search space is in many situations, quite large and there are complex interactions among parameters, namely if we are seeking to solve colour image segmentation problems. For instance, this has led [3] to adopt a GA to determine the parameter set that optimise the output of an existing segmentation algorithm under various conditions of image acquisition. That was the case for the optimisation of the *Phoenix* segmentation algorithm [22] described also in [4]. Another situation wherein GAs may be useful tools is illustrated in [23]. In [2] is defined a multi-term cost function, which is minimised using a GA-evolved edge configuration. Fuzzy GA fitness functions were also considered [7], mapping a region-based segmentation onto the binary string representing an individual, and

evolving a population of possible segmentations. Other implementations include the search of optimal descriptors to represent 3D structures [10], or the optimisation of parameters in GA hybrid systems [17]. Other authors [6] developed a GA based on a small set of manually-traced contours of the structure of interest (anatomical structures in 3D medical data sets). In [1] a GA is used to evolve a population of chromosomes that are distributed all over an environment. Because the segmentation progressively emerges as a by-product of a relaxation process [9] mainly driven by selection, the method has been called selectionist relaxation. In model designing terms, this last approach is indeed very close to that uses artificial ant colonies [20]. Approaches based on *Koza's* genetic programming paradigm (GP, [14]), i.e., GA used for finding appropriate algorithm structures and strategies, were also applied in image segmentation [18]. Finally, a fairly comprehensive review of other GA approaches in image processing is available in [5].

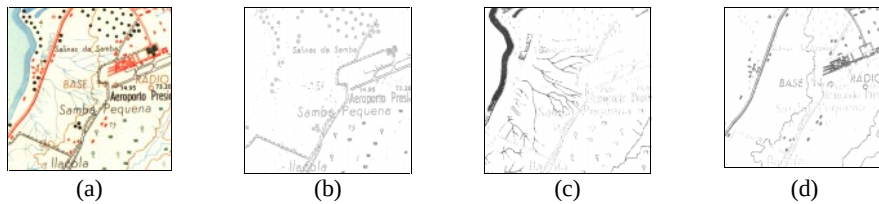


Fig. 1 - (a) Original Luanda (Angola) Colour Map [500x500 pixels] and (b) the respective 1st colour cluster segmenting train networks (in black), names (in black and dark brown), and parks (in dark green) pointed by the GA; (c) 2nd colour cluster - sea and rivers (blue); (d) Roads and buildings (red) and topological levels (brown). Other examples not shown here, reflect darker borders (red, black and brown) and background (whiter regions).

2 Genetic Clustering in Image Segmentation

As putted by *Andrey*, whether the GA is used to search in the parameter space of an existing segmentation algorithm [4], or in the space of candidate segmentations [2], an objective fitness function, assigning a score to each segmentation, has to be specified in both cases. However, evaluating a segmentation result is itself a difficult task. To date, no standard evaluation method prevails [25], and different measures may yield distinct rankings [24] (as an aside note, the present authors are nowadays developing image noise measures by mathematical morphology [21], allowing for instance, his use in image filtering design by GAs). One possible criterion is to think of homogeneous regions as the result of any appropriate and optimised clustering process, within the image feature space. Applications of GAs in clustering and grouping problems are intensively described in [12]. In the present approach, grey level intensities of RGB image channels are considered as feature vectors, and the *k*-mean clustering model (*J. MacQueen*, 1967) is then applied as a quantitative criterion (or GA objective fitness function), for guiding the evolutionary algorithm in his appropriate search. Since the *k*-mean clustering model allows to minimise the internal feature variance of each colour cluster (or the maximisation of the variances between different colour clusters [19]), natural and homogeneous clusters can emerge if the GA is properly coded. In other words, the image segmentation problem is simply reformulated as an unsupervised clustering problem, and genetic algorithms are then

used for finding the most appropriate and natural clusters. Since the clustering task can not be successfully applied within the image 2D space itself (e.g., similar pixels can be very far apart) the problem is coded within another space - that one of their colour features - 3D (grey level intensities, for the three channels). By this reformulation, one can in fact guarantee that similar pixels will belong to the same cluster. Preliminary efforts for this overall approach already were designed [19].

3 K-Means Clustering Model

K-Means clustering models were introduced in 1967 by *J. MacQueen*, and they are considered as an unsupervised classification technique. Once the method uses a minimum distance criteria, many authors consider this approach in many ways similar to the k-nearest neighbour rule method (there is also some similarities with the *Kohonen* LVQ method). In *MacQueen* terms however, k stands for the number of clusters searched by the model (and given as an input). All the strategy undergoes the minimisation of the expression (1). If one admits the partition of a p -dimensional space into c clusters (c colour classes), where n samples exists (n points characterised by p features), and being C_i each cluster i centre ($i = 1, \dots, c$), and X_j representative of each sample j co-ordinates ($j = 1, \dots, n$), where u_{ij} represents the hypothetical belonging of sample j into cluster i (i.e., $u_{ij} = 1$ if j belongs to cluster i ; $u_{ij} = 0$ if j belongs to any other cluster different from i). Naturally for the present case, the idea is to compute the colour cube partition minimising J by using genetic algorithms. We then have J as in Eq.1 ($u_{ij} = 0, 1$; $\sum u_{ij} = 1$ and $\sum \sum u_{ij} = n$):

$$u_{ij} \in U_{c \times n}; C_i = \frac{1}{n_i} \sum_{j=1}^n X_j; \min J = \sum_{i=1}^c \sum_{j=1}^n u_{ij} \|X_j - C_i\|^2 \quad (1)$$

4 Genetic Implementation

The above minimisation is then based on the different belonging combinations, of all points in the feature space. Naturally that, such task will be simply if the number of colours in one image to segment is low; however for high number of points in this 3D colour space (i.e., the different number of colours) this minimisation is hard to compute, since the combinatorial search space becomes very large. However, the partition of this 3D histogram into different clusters, must take in account the value of each point (that is, his frequency for a given RGB point). By this, the minimisation described in section 3, suffers a little modification (the method becomes weighted by this frequency f_j , since the number of colours of any RGB point are an important information in the overall process). The above J expression then becomes (Eq.2) $\min J = \sum \sum u_{ij} f_j \|X_j - C_i\|^2$. Another important issue in the GA implementation is the problem genetic coding. In order to do it appropriately, each chromosome codes the binary values of u_{ij} . However, to improve the GA search time and since the number of different colours in one colour image can be high, each 3D feature colour was submitted to a pre-partition. By this pre-procedure, the combinatorial search space is

reduced, as also as the number of bits in each GA chromosome. That is, each 3D colour cube (with side $256 - 8$ bit images) that could represent up to 256^3 colours, was reduced to a maximum of 512 points (i.e., 512 small cubes with side 32). In other words, all RGB points that fall into a small cube are agglomerated, being the new point represented by the centre of this small cube, and his frequency being equal to the sum of all frequencies of those points.

5 Results, Conclusions, and Future Work

The above strategy was applied in colour maps (214385 different colours-see figure 1). Since the colour map had roughly six type of colours the aim was to segment it into 6 clusters (6 prominent colours). The number of GA individuals were always 50, generations run = 10000, crossover mutation rate = 0.95, and mutation rate = 0.85. The respective computer time (PENTIUM 166 MHz / 32 Mb Ram) was 37.02 minutes. Finally, each string length (a parameter that depends only in the number of different colours present) was 468 bits long. Overall results points to highly satisfactory results, but with some problems in some clusters. The main reason is that the problem is by itself difficult (with large combinatorial search spaces), and that the pre-partition tends to reduce the discriminatory power of the overall strategy. This is mainly observed within pixels that form bounds of any important colour object. Image acquisition with low resolutions interpolates somehow their grey level intensities into intermediate values (between inner and outer bounds), and the result (with pre-partition) is significantly altered, since similar pixels can belong to different small cubes (naturally with low probability). However if the number of this kind of pixels is high, the strategy tends to create himself another cluster. On the other hand, and by observing the GA performance (J value) in each generation, we can conclude that similar results can be achieved with half of the generations run, since after 5000 generations J values are increasing very slow. Future work includes three main lines. First, to study the cluster relations (clouds of points) for each segmentation problem. This can bring useful information into the GA approach, and simply neighbourhood relations can be computed by using mathematical morphology on the 3D-colour cube. Second, more relevant segmentation measures must be studied. The present authors are making nowadays preliminary attempts, even if they are related with image noise [21]. Finally, significant improvements on the automatic design can be achieved by using ISODATA models - number of clusters can be automatically chosen by the hybrid search model.

References

- [1] Andrey P., 1999, "Selectionist Relaxation: Genetic Algorithms applied to Image Segmentation", *Image and Vision Computing* 17, 175-187.
- [2] Bhandarkar S.M., Zhang Y. and Potter W.D., 1994, "An Edge Detection Technique using Genetic Algorithm-based Optimisation", *Pattern Recognition* 27(9), 1159-1180.
- [3] Bhanu B., Lee S., 1994, *Genetic Learning for Adaptive Image Segmentation*, KAP.
- [4] Bhanu B., Lee, S. and Ming, J., 1995, "Adaptive Image Segmentation using a Genetic Algorithm", *IEEE Transactions on Systems, Man, and Cybernetics* 25(12), 1543-1567.

- [5] Bounsaythip C. and Alander J.T., 1997, "Genetic Algorithms in Image Processing - A Review", *Proc. of the 3rd Nordic Workshop on Genetic Algorithms and their Applications*, Metsatolo, Univ. of Helsinki, Finland, 173-192.
- [6] Cagnoni S., Dobrzeniecki A.B., Poli R. & Yanch J.C., 1999, Genetic Algorithm-based Interactive Segmentation of 3D Medical Images, *Image and Vision Computing* 17, 881-895.
- [7] Chun D.N. and Yang, H.S., 1996, "Robust Image Segmentation using Genetic Algorithm with a Fuzzy Measure", *Pattern Recognition* 29(7), 1195-1211.
- [8] Davis L.D., 1991, *Handbook of Genetic Algorithms*, Van Nostrand Reinhold, NYC.
- [9] Davis L.S. and Rosenfeld, A., 1981, "Cooperating Processes for Low-Level Vision: A Survey", *Artificial Intelligence* 17, 245-263.
- [10] Delibasis K., Undrill P.E., and Cameron G.G., 1997, "Genetic Algorithms applied to Fourier descriptor based Geometric Models for Anatomical Object Recognition in Medical Images", *Computer Vision and Image Understanding* 66(9), 286-300.
- [11] Duda R. & Hart P., 1973, *Pattern Classification and Scene Analysis*, J.Wiley & Sons, NYC.
- [12] Falkenauer E., 1998, *Genetic Algorithms and Grouping Problems*, J. Wiley & Sons, Boston.
- [13] Goldberg D.E., 1989, *Genetic Algorithms in Search, Optimisation and Machine Learning*, Addison-Wesley.
- [14] Koza J.R., 1992, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT Press, Cambridge, Massachusetts.
- [15] Michalewicz Z., 1996, *Genetic Algorithms + Data Structures = Evolution Programs*, 3rd Ed., Springer-Verlag.
- [16] Pal S. & Wang P. (eds), 1996, *Genetic Algorithms for Pattern Recognition*, CRC Press.
- [17] Poli R. and Valli G., 1993, "Neural Inhabitants of MR and Echo Images Segment Cardiac Structures", *IEEE Computer Society Press*, London, 193-196.
- [18] Poli R., 1996, "Genetic Programming for Image Analysis", in Koza, J.R., Goldberg, D.E., Fogel, D.B. and Riolo, R.L. (Eds.), *Genetic Programming 96, Proc. 1st Annual Conference*, Stanford Univ., MIT Press, 363-368.
- [19] Ramos V., 1997, *Evolution and Cognition in Image Analysis*, MSc Thesis (in Portuguese), 230 pp., Instituto Superior Técnico - IST, Lisbon, December 1997.
- [20] Ramos V. & Almeida F., 2000, "Artificial Ant Colonies in Digital Image Habitats - A *Mass Behaviour Effect Study on Pattern Recognition, accepted for ANTS'2000 - 2nd Int. Workshop on Ant Algorithms (From Ant Colonies to Artificial Ants), Brussels, Sep. 7-9.
- [21] Ramos, V. and Muge, F., 2000, "On Image Filtering, Noise and Morphological Size Intensity Diagrams", *Pattern Recognition Letters*, in press, (also on RecPad'2000).
- [22] Shafer S. & Kanade T., 1982, "Recursive Region Segmentation by Analysis of Histograms", *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 1166-1171.
- [23] Yoshimura M. and Oe S., 1999, "Evolutionary Segmentation of Texture Image using Genetic Algorithms towards Automatic Decision of Optimum Number of Segmentation Areas", *Pattern Recognition* 32, 2041-2054.
- [24] Zhang Y.J. & Gerbrands, J.J., 1994, "Objective and Quantitative Segmentation Evaluation and Comparison", *Signal Processing* 39, 43-54.
- [25] Zhang Y.J., 1996, "A Survey on Evaluating Methods for Image Segmentation", *Pattern Recognition* 29(8), 1335-1246.

Spoken Query Processing for Information Access in Digital Libraries

Fabio Crestani

Department of Computer Science
University of Strathclyde
Glasgow G1 1XH, Scotland
`fabioc@cs.strath.ac.uk`

Abstract We briefly outline the ongoing research at Strathclyde University on the use of spoken query processing for information access in digital libraries.

1 Spoken Access to Digital Libraries

The widespread access to Internet and the high bandwidth often available to users makes them assume that there is nothing easier than remotely connecting to and searching a large information repository. Nevertheless, a very large part of the world population does not and will not have for a long future access to computers or Internet. There are cases in which the only available (or the most convenient) communication mean is a telephone or mobile phone. In addition, there are also situations and users for whom the availability of computer screen and keyboard is not useful. Blind or partially-sighted users (e.g. users who have problems due to disabilities, protective clothing or working environment) may only have access to information if this is accessible via voice and/or sound. In all these cases, if we want users to take advantage of the large amount of information stored in digital repositories, it is necessary to enable access to them solely via voice. This implies the design and implementation of systems capable not only of understanding the user's spoken request, finding the required information and presenting it as speech, but also capable of interacting with the user in order to better understand his/her information need, whenever this is not clear enough to proceed effectively with the searching.

At Strathclyde University we are investigating how we can improve current information access technology to enable user to access digital libraries using voice. At the current stage, the main objective of the project is to enable a user to interact with a probabilistic Information Retrieval (IR) system over a telephone line. This implies enabling the user to submit queries, commands and relevance assessments via voice and be presented with the content of retrieved documents using an interactive vocal interface. A first prototype system enabling journalist to access newspapers archives using a telephone or mobile phone, is currently being implemented and tested [2]. In this short paper we outline some of the ongoing research on improving the effectiveness of spoken query processing for information access in digital libraries.

2 Spoken Query Processing

While the effectiveness of spoken document retrieval (SDR) has been well studied and well documented in recent IR literature, very little research work has been devoted to studying the effectiveness of spoken query processing (SQP). It has to be recognised that SQP poses a number of additional challenges compared with SDR. The most important ones are:

- query processing needs to be performed on-line and “almost” real time, while spoken document recognition and indexing can be performed off-line;
- queries are usually much shorter than documents and word recognition errors (WRE) may have more serious effects on them;
- we may have very little training data on the voice of each user and we may have a large number of different users in different acoustic conditions.

Given these challenges, we decided that our first priority was to test the limitations of current information access technology in the SQP task, given realistic levels of WRE rates.

3 Effects of Word Recognition Errors in SQP

In [4] we reported the results of an experimental study of the effects of WRE in spoken queries on the effectiveness of an IR system. The document collection we used was generated at Dragon Systems Inc. [1], and this is currently the only test collection available with spoken queries. A large number of experiments using a standard IR system were carried out. In these experiments some of the parameters of the IR process (i.e. stemming, stoplist, weighting schema, etc.) were changed to study their effects on the effectiveness on the IR task in relation to the different levels of WRE rates. For reason of space we cannot report here the results obtained, but we can say that the results show that classical IR techniques are quite robust to considerably high levels of WRE rates (up to about 40%), in particular for long queries.

However, our experimentation falls short in a number of ways. First, the queries used were too long and not really representative of typical user queries (although some initial unpublished user studies on spoken queries indicate that spoken queries are usually longer than written queries). Second, the WRE rates of the queries used in this experimentation were typical of “dictated” spoken queries, since this was the way they were generated. Dictated speech is considerably different from spontaneous speech and easier to recognise. Spontaneous spoken queries have higher levels of WRE rates and different kinds of errors. Finally, the queries used were generated artificially from queries spoken in a laboratory environment. Telephone speech is more difficult to recognise than laboratory speech and produces higher WRE rates.

Initial experiments with a more realistic test collection show that standard IR techniques are not so robust to the kind of WRE found in spontaneous speech. More complex techniques for dealing with SQP of spontaneous speech are therefore needed.

4 Techniques for Improving SQP

We are currently experimenting with a number of techniques for obtaining acceptable levels of effectiveness of SQP of spontaneous speech for information access. Here, for reasons of space, we can only outline the first results obtained and the directions of the work in progress.

Relevance Feedback. Given the level of IR effectiveness with spoken queries obtained in [4], we can conclude that it will be quite likely that in the first n retrieved documents (with n dependent on the user's preference and usually less than 10) there will be at least one relevant document. In a previous study we have successfully tested the ability of the user to perceive if a document is relevant to an information need when the document is presented in the form of a short spoken summary [8]. Such scenario is consistent with our current target prototype application. The above two results enabled us to conclude that *relevance feedback* (RF) could be a good strategy to improve effectiveness in a SQP task.

Indeed, in [6] we showed that both standard RF and pseudo RF enable to improve the IR effectiveness with spoken queries, in particular for short queries. In fact, the use of RF (and in particular of standard RF) enables to reduce considerably the difference in effectiveness between long and short queries previously found in [4].

Use of Prosodic Stress for Topic Detection. In [7] we investigated the relationship between acoustic stress in spoken sentences and information content. On one side, the *average acoustic stress* was measured for each word throughout each utterance. On the other side an IR index (the *tf-idf* weight) was calculated. The scatter plots of the two measures showed a correlation between higher values of average acoustic stress with increasing values of the IR index of the word in the majority of the analysed utterances. A statistically more valid proof of such a relationship was derived from the histogram of the words with high average acoustic stress vs. the IR index. This confirmed that a word with high average acoustic stress has also a high value of the IR index and, if we trust IR indexes, also high informative content.

This study confirmed our hypothesis of a direct relationship between acoustic stress and information content as identified by IR weighting of spontaneous spoken sentences. The next stage of this work will be the integration of prosodic stress and IR weighting evidence into a new IR weighting algorithm for spontaneous sentences. This weighting schema will take into account both word acoustic and statistical clues to characterise the document/query informative content. It will be extremely useful in a number of tasks and in particular for SQP where the short length of queries requires every possible clue to fully capture the user's information need.

Combination of Semantic and Phonetic Term Similarity. A fundamental problem of IR is “term mismatch”. A query is usually a short and incomplete description of the user’s information need. Users and authors of documents often use different terms to refer to the same concepts. This fact produces an incorrect relevance ranking of documents with regards to the information need expressed in the query [5]. A similar problem can be found in SDR and SQP, where terms misrecognised by the speech recognition process are found not matching in query and document representations. Naturally, this hinders the effectiveness of the IR system. We called this problem “term misrecognition”, by analogy to the term mismatch problem.

In [3] we presented a model for dealing with the term mismatch and the term misrecognition problems in SDR and SQP. Term similarity is used at retrieval time to estimate the relevance of a document in response to a query by looking not only at matching terms, but also at non-matching terms whose semantic and/or *phonetic similarity* are above a predefined threshold. Semantic similarity can help solve the term mismatch problem. It can be estimated using Expected Mutual Information Measure or some similar measure. Phonetic similarity, on the other hand, can help tackle the term misrecognition problem. It can be estimated using Error Recognition Confusion Matrices, for example. An experimental investigation is currently being carried out. The experimental results will provide useful feedback on the effectiveness of the models proposed and on how to effectively combine semantic and phonetic similarity.

References

1. J. Barnett, S. Anderson, J. Broglio, M. Singh, R. Hudson, and S.W. Kuo. Experiments in spoken queries for document retrieval. In *Eurospeech 97*, volume 3, pages 1323–1326, Rhodes, Greece, September 1997.
2. F. Crestani. Vocal access to a newspaper archive: design issues and preliminary investigation. In *Proceedings of ACM Digital Libraries*, pages 59–68, Berkeley, CA, USA, August 1999.
3. F. Crestani. Combination of semantic and phonetic term similarity for spoken document retrieval and spoken query processing. In *Proceedings of the 8th Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, Madrid, Spain, July 2000. In press.
4. F. Crestani. Effects of word recognition errors in spoken query processing. In *Proceedings of the IEEE ADL 2000 Conference*, Washington DC, USA, May 2000. In press.
5. F. Crestani. Exploiting the similarity of non-matching terms at retrieval time. *Journal of Information Retrieval*, 2(1):23–43, 2000.
6. F. Crestani. Word recognition errors and relevance feedback in spoken query processing. In *Proceedings of Fourth International Conference on Flexible Query Answering Systems*, Warsaw, Poland, October 2000. In press.
7. R. Silipo and F. Crestani. Acoustic stress and topic detection in American English spoken sentences. Technical Report TR-00-005, International Computer Science Institute, Berkeley, CA, USA, February 2000.
8. A. Tombros and F. Crestani. Users’s perception of relevance of spoken documents. *Journal of the American Society for Information Science*, 2000. In press.

A Metadata Model for Historical Documentary Films

Giuseppe Amato, Donatella Castelli, and Serena Pisani

Istituto di Elaborazione dell'Informazione – CNR – Pisa (Italy)
{amato,castelli,serena}@iei.pi.cnr.it

Abstract. This paper presents a metadata model for historical audio-video material able to describe information for supporting both the traditional archival functions and the advanced applications, like video summary, speech recognition, and automatic semantic content extraction.

1 The Metadata Model

The metadata model presented in this paper originated from our experience within the European Chronicles On-line (ECHO) project¹. The aim of this project is to build a digital library to permit cross-archives searching and retrieval of historical documentary films by exploiting very advanced audio-visual technologies.

In the ECHO digital library, historical documents regarding the war and post war period are considered; in particular, *newsreels*, *documentaries*, *reportages* and *cuts* are collected. Newsreels are the parents of today's news television, that were broadcast at cinema before the beginning of a film and the typical language used is propaganda language. Documentaries are more detailed and longer films mainly focused on one particular aspect. Reportages are subparts of several newsreels. Cuts corresponds to the video material that was collected, but never broadcast.

The metadata model is depicted in Figure 1. It resulted from the analysis of the answers to three different questionnaires tailored for the historical documentary film archivists, the technology providers and the speech recognition and digital video abstracting researchers. This analysis led us to conclude that a multi-layered and hierarchical description of audio-video documents was needed. The former characteristic allows describing different aspects of the same document. The latter permits to support both (i) the interoperability among different archives at the corresponding hierarchical levels and (ii) to satisfy the needs of special interest user communities. This last point is simply obtained by refining the model by adding specialized descriptions.

This multi-layered and hierarchical structuring is in accordance with the IFLA model [3]. IFLA defines the four entities WORK, EXPRESSION, MANIFESTATION and ITEM describing different aspects of the resource.

In this paper we present the structure of the proposed metadata model and omit the metadata fields associated with each entity. They can be found in [1]. In our model each of the four IFLA entities has been considered as the most general entity of the

¹ ECHO - Project number 11994, Information Societies Technology (IST) Programme, funded by the European Commission.

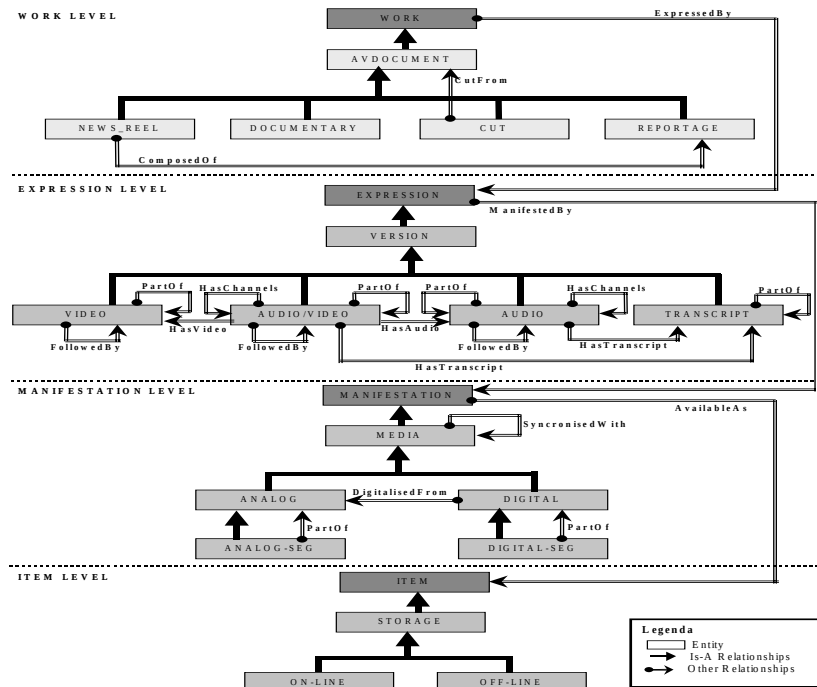


Fig. 1. The metadata model structure

corresponding level. Each entity has then been specialised to perform the structure of the metadata fields. At the Work Level the entity AVDOCUMENT is a generic entity representing an intellectual or artistic film work, while its subentities NEWSREEL, REPORTAGE, CUT, and DOCUMENTARY represent the types of historical audio-video documentary films. A newsreel may be composed by a sequence of reportages that correspond to various parts of the whole audio-video document itself. This situation is modelled by the relationship *ComposedOf*. This structure emphasizes the fact that some reportages are used as part of different documentary films. The elements in AVDOCUMENT may be related to the cuts from which are extrapolated through the relationships *CutFrom*.

Each document, belonging to the entities of the Work Level, may have one or more versions. For instance, an element of VERSION can be the audio realization of a newsreel, another the audio-video realization of the same newsreel. A decomposition of VERSION in VIDEO, AUDIO, AUDIO/VIDEO and TRANSCRIPT was introduced to endow these particular kind of versions of their own set of metadata fields.

At the Manifestation Level, the entity MEDIA represents different supports that can be used to maintain documents. The sub-entities corresponding to segments are ANALOG-SEG and DIGITAL-SEG. The sub-entities ANALOG and DIGITAL represent, instead, entire documents.

The entity STORAGE at the Item level describes general information about individual copies of the same document. Specific information about on-line and off-line

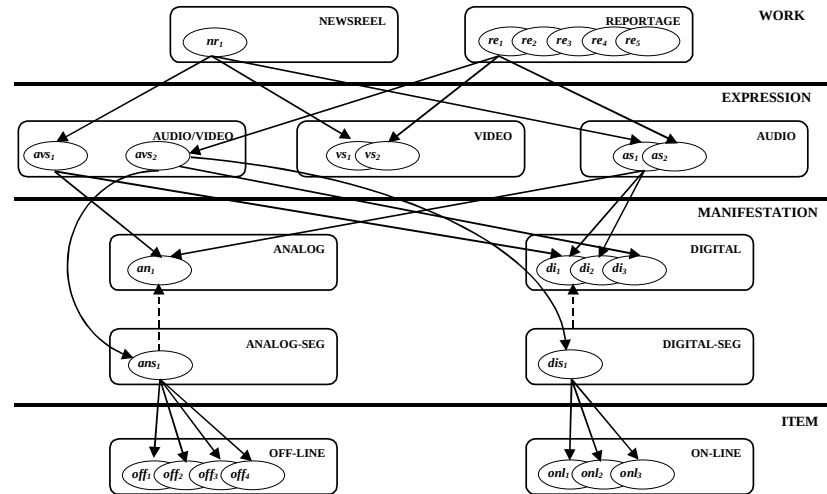


Fig. 2. A running example

document copies is, respectively, modeled by the sub-entities ON-LINE and OFF-LINE. This partition permits, for instance, to represent a situation in which the same MPEG1 file may be stored on servers with different load and access times, or in which several copies of the same CDROM or of the same VHS tape have different quality of preservation.

2 A Running Example

Figure 2 outlines this example. In the figure, boxes correspond to entities, ellipses to objects, arrows highlight relationships between objects at different levels, dashed arrows represent part-of relationships.

Let's suppose that we have a newsreel about the Landing of the allied forces in Normandy. This newsreel is represented by an object nr_1 of the NEWSREEL entity of the work level. The "Landing in Normandy" newsreel is composed of several reportages. These components are objects of the REPORTAGE entity. We suppose that there is the object re_1 representing the reportage where the US president Roosevelt announces the attack to the press; the object re_2 that presents the reaction of Churchill; the objects re_3 , re_4 , re_5 , that are about various actions taken directly during the attack.

Let's suppose that our newsreel is available in different versions. Different objects, at the expression level, represent these versions. These versions may belong to different entities, or may be different objects of the same entity. For instance, the object avs_1 , of the AUDIO/VIDEO entity represents the Italian spoken version of the newsreel (this is obtained by inserting the value "Italian" in the field language of these objects); the object vs_1 , of the VIDEO entity, is a video only expression; the object as_1 , of the AUDIO entity, is an audio expression, corresponding to a radio program announcing the attack. Of course, reportage objects may also have several expressions. For instance we suppose that re_1 , is represented by three expressions.

Let's suppose that the audio/video expression avs_1 is available as VHS videotape and also as a MPEG file. This is modeled by two objects at the manifestation level: the object an_1 of the ANALOG entity and by the object di_1 of the DIGITAL entity. The audio expression as_1 is embodied both in the objects an_1 and di_1 , so it can also be represented by these objects. In addition, we suppose that there is a separate MP3 file that only contains the audio of the newsreel. The di_2 object of the DIGITAL entity may represent it. Let's now consider the audio/video sequence avs_2 corresponding to the reportage with President Roosevelt. We suppose that there is a QuickTime file that contains it. The file is represented by the object di_3 of the DIGITAL entity. In addition, since it is also contained in the whole newsreel, there are two objects ans_1 and dis_1 , respectively of the ANALOG-SEG and DIGITAL-SEG, that represent the portion of the objects an_1 and di_1 , corresponding to VHS and MPEG supports of the entire newsreel, where the reportage appears.

Let's now suppose that the library has four copies of the VHS videotape an_1 that contains our newsreel. These copies are associated, at the item level, with the objects $offl_1$, $offl_2$, $offl_3$, and $offl_4$ of the OFF-LINE entity. Continuing with the example, we can suppose that the MPEG file represented by di_1 is mirrored on three different servers. These files are then associated respectively with the objects onl_1 , onl_2 , and onl_3 of the ON-LINE entity.

3 Concluding Remarks

This paper introduced a metadata model for historical documentary films. This model is the result of an interaction with the data, technology and services providers involved in the ECHO project. This interaction outlined two realities: (i) the reality of the traditional archivists which were accustomed to maintain information about the content of the video, its location and its cataloguing process; (ii) the reality of the technology providers that describe the video mainly in terms of its physical characteristics. Most of the existing metadata models, taking into account audio-visual resources, tend to focus on one of these two realities. The model presented is an attempt to cover both of them. It differs from other similar attempts [2] for its approach towards interoperability. The metadata fields, representing different aspects of the audio-visual resource, have been described at different levels of abstraction. This structure provides a great potential for interoperability. We are currently specifying the model as a simple RDF schema [4].

As concluding remarks, let us outline that even if this work focused on the historical documentary films, we are confident that our approach may be generalised to define models for other kind of audio-visual resources.

References

- [1] The ECHO Metadata Model. G. Amato, D. Castelli, S. Pisani, P. Venerosi. IEI-CNR Technical Report, 2000.
- [2] A Comparison of Schemas for Video Metadata Representation. J. Hunter, L. Armstrong. WWW8, Toronto, May 10-14, 1999.
- [3] Functional Requirements for Bibliographic Records. Final Report. K. G. Saur München, 1998. <http://www.ifla.org/VII/s13/frbr/frbr.pdf>
- [4] Resource Description Framework (RDF) Home Page: <http://www.w3.org/RDF/>

Image Description and Retrieval Using MPEG-7 Shape Descriptors

Carla Zibreira and Fernando Pereira

Instituto Superior Técnico/Instituto de Telecomunicações, Av. Rovisco Pais, Lisboa, Portugal
Carla.Zibreira@lx.it.pt, Fernando.Pereira@lx.it.pt

Abstract. The increasing amount of digital audiovisual information and the need to efficiently and effectively describe and retrieve this information as well as the big technological developments in the related domains have been acknowledged by MPEG (Moving Pictures Experts Group) by initiating a new work item, formally called “Multimedia Content Description Interface” but better known as MPEG-7. This paper will introduce the MPEG-7 standard, with special emphasis on the adopted shape descriptors: Curvature Scale Space and Zernike Moments. Finally, the description and retrieval mechanism based on the MPEG-7 shape descriptors developed will be presented.

The Context

Digital audiovisual information is nowadays accessible to everybody, not only in terms of consumption but increasingly also in terms of production. But if it is today easier and easier to acquire, process and distribute audiovisual content, it must also be equally easy to access the available information because huge amounts of digital audiovisual information are being generated, all over the world, every day. Unfortunately, identifying the desired information, by retrieving or filtering, is becoming more and more difficult.

Due to the increasing need of automatically retrieving audiovisual information, since there are ever more activities where this kind of information is required, various description and retrieval mechanisms have emerged in the Internet although most of them still based on textual descriptions. Consequently, much of the audiovisual information found over the web is found by pure luck or difficultly because of the enormous number of irrelevant results to each query made. Even though these mechanisms are not the best solution, the majority of the users still keeps using them in spite of their limitations, which in fact highlights how big is the need that they address and thus the urge to solve the problem of audiovisual content description in a more efficient and adequate way.

Nowadays, much of the audiovisual content available is being manually annotated, which obviously is no more an acceptable method for two major reasons: the exponentially growing amount of information to describe would imply great costs and human resources, and the subjectivity inherent to textual descriptions would restrict the usage of the content in different application domains. These facts highlight the need to describe automatically and objectively the audiovisual information, requiring tools that automatically extract more or less complex audiovisual features from the content and substitute or complement manual descriptions. Unlike the textual descriptions, these audiovisual features will allow the content to be described advantageously in three ways: automatically in which case not the specialists but the machinery will worry about the great amount of information to describe, objectively

where problems such as subjectivity and specialization are eliminated, and adapted to the audiovisual content (it is audiovisual not textual) allowing queries to be formulated in a way more adequate to the content in question, e.g. using colors, shapes and motion.

The MPEG-7 Standard

The growing amount of on-line information, the consequent increasing need to efficiently and effectively retrieve audiovisual information and the big technological developments in the related domains have been acknowledged by MPEG (Moving Pictures Experts Group) [1], which is the ISO standardization committee responsible for important audiovisual representation standards such as MPEG-1, -2 and -4. In this context, MPEG initiated a new work item, formally called “Multimedia Content Description Interface” but better known as MPEG-7, with the objective to specify a standard way of describing various types of audiovisual information, including images, video, speech, audio, graphics, 3D models, and synthetic audio, irrespective of its representation format, e.g. analog or digital, and storage support, e.g. paper, film or tape [2].

Since the MPEG-7 emphasis is on audiovisual content but textual information is still of great importance in the description of audiovisual information, the standard will not specify new description tools for text, but will rather use existing solutions for describing text documents, and will support them as appropriate [3]. Moreover, MPEG-7 will allow linking audiovisual data descriptions to any associated data, notably textual data. MPEG-7 will be generic, and not especially tuned to any specific application. Like the other members of the MPEG family, MPEG-7 will be a standard representation of audiovisual information standardizing the minimum number of tools in order to guarantee interoperability. To address the audiovisual description challenge, MPEG-7 will standardize four types of tools [3]:

- **Descriptors** – define the syntax and semantics of a representation entity for a feature, e.g. time-code for representing duration, color moments and histograms for representing color, a character string for representing a title.
- **Description Schemes** – consist of one or more descriptors and description schemes as well as the structure and semantics of the relationships between them, e.g. a movie, temporally structured as scenes and shots, including some textual descriptors at the scene level, motion descriptors at the shot level, and color and shape descriptors for the key-frames in each shot.
- **Coding Schemes** – allow to generate descriptions encoded to fulfill relevant requirements, such as compression efficiency, error robustness, and random access.
- **Description Definition Language (DDL)** – allows to specify and extend the existing description schemes and possibly descriptors, as well as the creation of new ones.

These four types of tools are the normative elements of the standard, implying that they will have to be implemented as defined by the standard in order to guarantee the interoperability of the description and retrieval mechanisms. Consequently, and besides its fundamental importance, the feature extraction tools, the querying methods, the similarity measures and the description generation and coding processes are not standardized because this is not essential to guarantee interoperability between different mechanisms. This strategy ensures that good use can be made of the

continuous improvements in the relevant technical areas with the consequence that new automatic analysis tools can always be used, also after the standard is finalized, and that it is possible to rely on competition for obtaining ever better results. The description of content may be done using the so-called low-level features, such as color and shape for images, and pitch for speech, as well as through higher-level features like genre classification, and rating, with a semantic value associated to what the content means to humans. Although low-level features are easier to extract since they are automatic and objective (they can typically be extracted fully automatically), the truth is that most (non-professional) consumers would like to express their queries as much as possible at the semantic level (which is rather difficult using only automatic extraction). As it is possible to design a description framework that combines low-level and high-level features, there is no need to choose one and only one of the two types of features. This low-level versus high-level features dichotomy is central to MPEG-7 and constitutes one of the major differences between MPEG-7 and other emerging audiovisual description solutions. In fact, description genericity requires low-level, non-context dependent, features, pointing towards the harmonious integration of low-level, signal processing based, features and high-level, audiovisual understanding based, features which convey semantic meaning.

Following the interactivity trend which states that users want to access deeper than the frame level in terms of video content, MPEG-7 will allow to independently describe not only video scenes but also the various objects in the scene adopting the object-based data model already adopted by the MPEG-4 standard [4]. Adopting the object-based data model means that the low-level description of each video object in the scene has to consider three major types of data: texture, motion and shape. MPEG-7 has already adopted a set of descriptors to represent relevant features associated to these data [5]. These features can be automatically extracted to objectively describe the content without any type of semantic, application dependent, mapping. This will allow the same objective description to be used in many different semantic environments, by having the low-level – high-level semantic mapping made at the time of the retrieval. This will give power to an essential rule: “a content asset is everything it can be and not only what an indexing expert at a certain moment decided it to be”.



Fig. 1. Image and corresponding shape information: a) bream, b) children and c) container

An Application for MPEG-7 Shape Description and Retrieval

The object-based data model adopted by the MPEG-4 and MPEG-7 standards brought for the first time to the international standardization arena the shape information. Shape information is essential both for the coding as well as for the description of arbitrarily shaped objects. While MPEG-4 defined the first shape coding standard, it is the task of MPEG-7 to define the first set of standard shape descriptors. For the moment, MPEG-7 already chose two shape descriptors: a contour-based descriptor (description of the closed contour of a simple object or region) based on the Curvature Scale Space (CSS) representation and a region-based descriptor (description of

complex objects defined as a set of regions) based on Zernike moments [5]. These descriptors shall provide in an efficient way all the relevant shape related functionalities, whatever the application domain.

This paper presents an MPEG-7 based image description and retrieval application for shape information, developed at Instituto Superior Técnico. This application allows the current MPEG-7 shape description tools to be evaluated and compared regarding some alternative description tools, notably for shape information. Besides the standard MPEG-7 shape descriptors, Multi-Layer Eigen-Vectors and turning angles shape descriptors are already been implemented. All the tools have been implemented in the context of the MPEG-7 Experimentation Model (XM) software which allows the application to exploit all the other description tools available within MPEG-7.

In terms of description, the application allows the integration of several low-level descriptors, e.g. shape and colour, in the same description. The retrieval mechanism uses a dual solution for shape information based on retrieval by example and by sketch. Moreover several similarity measures are being studied with the implemented shape descriptors in order to conclude about the best matching procedure for shape information.

The application developed allowed the comparative performance analysis of the implemented shape descriptors (standard and non-standard) using the MPEG-7 data set and well defined experimental conditions (using MPEG-7's evaluation methods). Future work will consider the development of high-level semantic mapping methods based on low-level shape descriptors to allow more intuitive image and video retrieval procedures.

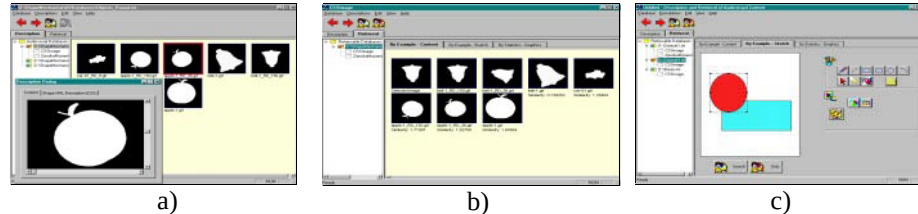


Fig. 2. Application: a) Description Interface, Retrieval Interface b) by Example and c) by Sketch

In conclusion, this paper proposes an application allowing to study and evaluate a new type of tools – *shape description tools*. These tools support new functionalities in the context of the new object-based visual data representation model more adapted to provide the object-based retrieval, manipulation, interaction and customisation capabilities more and more required by multimedia applications and users.

References

- [1] MPEG Home Page, <http://www.cselt.it/mpeg/>
- [2] MPEG Requirements Group, "MPEG-7: Overview", Doc. ISO/MPEG N3445, MPEG Geneva Meeting, June 2000.
- [3] MPEG Requirements Group, "MPEG-7 Requirements Document", Doc. ISO/MPEG N3446, MPEG Geneva Meeting, June 2000.
- [4] F. Pereira, "MPEG-4: Why, What, How and When?", Image Communication Journal, vol. 15, n° 4-5, December 1999.
- [5] MPEG Video Group, "MPEG-7 Visual Part of XM", Doc. ISO/MPEG N3398, MPEG Geneva Meeting, June 2000.

A Large Scale Component-Based Multi-media Digital Library System

– Development Experience and User Evaluation –

Hiroshi Mukaiyama

Japan Information Processing Development Center (JIPDEC)
Kikai-Shinko Kaikan Building, 3-5-8, Shibakoen, Minato-ku, Tokyo, Japan
hiro@jipdec.or.jp

Abstract. The Next Generation Digital Library Project has developed a 3-tire client/server system used CORBA and an agent technology, aiming a component-based digital library development. We invented the new functions such as an intellectual property right management and a bill charge in the digital library's functions and conducted a user evaluation test. This paper presents these technologies and the evaluation results.

1 Introduction

An advent of the Internet and WWW has enabled new knowledge accumulation and sharing way so called digital library (DL). Since current Internet and WWW are basic technologies, further researches and developments are needed to develop the large scale DLs and the information sharing systems hereafter.

The Ministry of International Trade and Industry (MITI) decided to launch new R & D named the Next Generation Digital Library Project [1, 2].

On the premise that the future systems would use the 3-tire client/server structure, we decided to develop each technology in a way that fits in the 3-tire model and adopted OMG's CORBA (Common Object Request Broker Architecture) as a component object model.

This paper describes the developed components in section 2 and a prototype development in section 3. Followed is a user evaluation in section 4 and we conclude in section 5.

2 Software Components

The component development is a first step of CBSD (Component-Based System Development) that we tried. The developed components vary from the middlewares that become an infrastructure for the systems, to the application-level ones. Here we describe only part of the components that are incorporated in the prototype system explained in section 3.

Object oriented RDB

Since RDBs are functionally insufficient for multi-media data, we made object-oriented expansion based upon SQL99 that we can define new class like SGML class and Image class. We created a mechanism that the methods are implemented as plug-ins.

Agent management middleware

Considering that agents are objects that hold autonomy and mobility, we developed an agent life cycle management system and an agent communication mechanism.

Document management middleware

We developed a document management middleware that implemented AIIM's DMA model as a base model and extended the model for DL's use.

Video management middleware

We developed a video management middleware that can handle DBs and media servers consistently. We also developed a video meta data description language, which is called VCML (Video Content Markup Language) that enables partial retrieval of the video.

Simultaneous search

This is a function to retrieve external DLs simultaneously. We developed this function as a multi-agent system used a contract-net model. By the bidding process of contract-net, recommendation of the best DLs and reduction of access cost become possible.

Information recommendation

This is a technology to connect contents providers and their users. We developed a new contents recommendation method that uses CBF (contents-based filtering) and SIF (Social Information Filtering) in a complementary way.

Information Filtering

We developed a method to create new filter by combining the filters like camera lens.

Automatic classification

We developed a new mechanism to classify the contents automatically calculating relations between the terms in the classification and the terms in the contents.

Intellectual property right management

We developed an IP management system integrated the user certification, ciphering, digital watermark and event-disable technology. Since it is impossible to protect IP perfectly by software we set a goal that could be accepted in a community.

Content entry framework

We developed an open framework that can insert and change the functional parts such as image scanning, layout analysis, character recognition and tagging.

3 System Design and Fabrication

The DL development that is a second step in CBSD begins by matching the components and the user requirements. We started the prototype development by getting cooperation from the National Diet Library (NDL) who is the largest library in Japan and has a plan to develop its own DL. In the phase of requirement specification we explained our developed components to staff of NDL at first. After that, we designed an overall system based on the selected components. In the area that is not fulfilled by our components, wrapping of legacy technologies and the supplemental development of functions were taken. User interface and bill charge are those ones.

Fig.1 illustrates the system configuration. As to CBSD, there are reports that it caused an architecture mismatch such as event handling. However there was no big problem for our case thanks to CORBA.

Although CORBA promises a location transparency of objects, this notion caused sometimes performance problems. We should care the platform that objects can work.

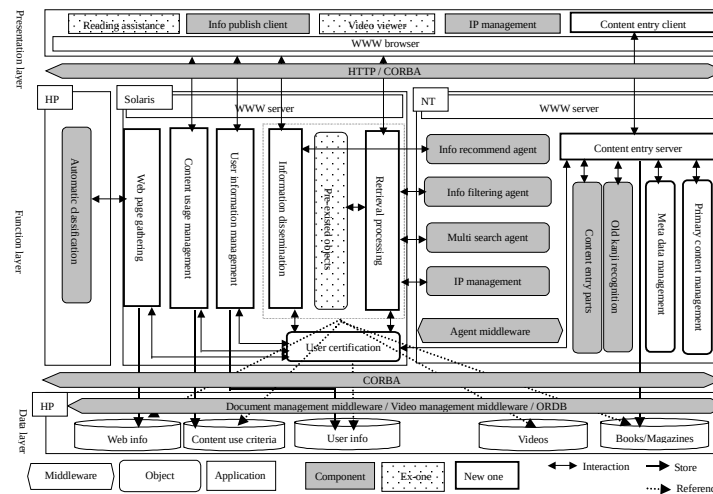


Fig. 1. System configuration

4 User Evaluation

After completion of the development, we conducted the user test for about one month whose ninety-three examiners were selected mostly from staff of many kinds of

libraries to evaluate component effectiveness and system effectiveness. We prepared 100,000 bibliographic data, 300 full text materials, 8,500 page images, 800 magazine articles, 8,000 web home pages, 40 hours news videos and 34 hours technical videos. As overall evaluation, we got positive answer that is effective from more than 70 percent examiners. Amid vague image of future digital library system we received a lot of response that our system is very informative for their future systems.

The items that got good score were: unified retrieval for multi-resources; stepwise retrieval expansion from bibliography to index, to abstract, to contents; video retrieval by scene; cross search for external libraries; bill charge for various unit such as book, article and figure. On the other hand, the items needed to improve were: combination of index search and full text search; recommendation criteria; feedback way of filtering.

5 Conclusions

We think that our results are evaluated from two points that are the system construction method and the new functions for DLs.

As for the system construction, we showed that new development methodology was feasible which develops the components based on CORBA at first and fabricates the system from the components. And we realized that the middlewares like the document management middleware and the video management middleware are very effective as a common infrastructure and accelerate the development.

Furthermore we invented the new system configuration that objects and agents can coexist and communicate each other through the development of agent-based services like the cross search and the information recommendation. This means that a technical base was created to move from object-orientation to agent-orientation or to mix them.

As for the DL's functions, we developed the new features such as intellectual property right management and billing that are rarely deployed at current DLs and conducted user evaluation test. As a result we did not get serious response and we assume that our method is accepted as a proper way. This might be good example for future DL development. However this does not mean that future DLs should have billing in mind. The bill charge in the library is still open issue in Japan.

Acknowledgements:

The Next Generation Digital Library Project has been carried out under the MITI/IPA (Information-technology Promotion Agency)/JIPDEC (Japan Information Processing Development Center) with Hitachi, Fujitsu, NEC, IBM, Toshiba, Mitsubishi Electric, Oki Electric Industry, Unisys.

I would like to thank the members of project and contributed persons.

References

1. Hiroshi Mukaiyama: Technical Aspect of Next Generation Digital Library Project, ISDL'97, pp.72-79, 1997.
2. JIPDEC ed.: Proceedings of Next Generation Digital Library Project, March, 2000 (in Japanese).

Personalised Delivery of News Articles from Multiple Sources

Gareth J. F. Jones, David J. Quested, and Katherine E. Thomson

Department of Computer Science, University of Exeter, EX4 4PT, U.K.

email: G.J.F.Jones@exeter.ac.uk,

WWW home page: <http://www.dcs.ex.ac.uk/~gareth>

Abstract. Traditional news media report a single set of articles on current news stories. Online news sources make multiple stories on the same topic available reflecting different perspectives on the same news event. Navigating between these news sources to find stories of interest can be time consuming and inefficient. These multiple stories can be combined into personalised news packages by selecting items on topics of interest to an individual user. The appropriate contents of these personalised news packages can be determined by a combination of information retrieval techniques and explicit user preferences. This paper describes systems exploring this approach to personalised news delivery.

1 Introduction

News has traditionally been delivered in pre-packaged forms originally in newspapers, which have subsequently been joined by radio and television broadcasts, and most recently by internet news services. In the conventional model these news packages have a number of common features: the contents are decided by the editor and the contributing writers, the coverage of stories represents a national or sometimes regional perspective, and the depth of coverage of an individual story is determined by the editors' judgment of the general readership's interest in it. The ongoing expansion in the availability of electronic news material provides immediate access to many different perspectives on the same news stories. From these services a user can find articles on topics that interest them personally rather than those chosen by the editor of a particular news provider. The main problem for a user attempting to exploit this wide range of options is how to determine which articles to access without spending inordinate amounts of time engaged in web searching.

This paper introduces our work on the development of integrated personalised news delivery services from multiple news sources. Our systems seek to combine the convenience of existing news media by taking the format of a traditional newspaper as a metaphor with which users are familiar, but including stories taken from multiple news sources which are likely to be of interest to an individual user. The systems can link related stories so that the user is easily able to follow different personal or international perspectives on the same topic. The user is given full control of the information sources to be used.

The remainder of this paper describes the features of differing news media and how these can be addressed in an online personalised news delivery system,

outlines the topic and story selection methods used, introduces the user interface approach of our current techniques, discusses the limitations of current web technology, and considers directions for further work.

2 Existing News Delivery Systems

While sharing many common features, existing news media sources adopt different models for updating their content. Paper newspapers are usually published daily with the obvious, but important, consequence that their content can only be updated daily. Radio and television news broadcasts appear with a variety of frequencies ranging from several times a day to updates several times an hour. Finally internet news sources are usually updated on an ad hoc basis as important news is received. A personalised news delivery system can adopt components from any of these news selection strategies. The rate at which the content is changed can be determined by the user based on the sources selected.

3 Personalised News Delivery Systems

There are clearly many options in the design of personalised news delivery systems. We have recently developed several systems to experiment with source selection, user interaction and presentation strategies. Although these systems vary in appearance and detailed functionality, they have the same underlying philosophy as outlined in this section.

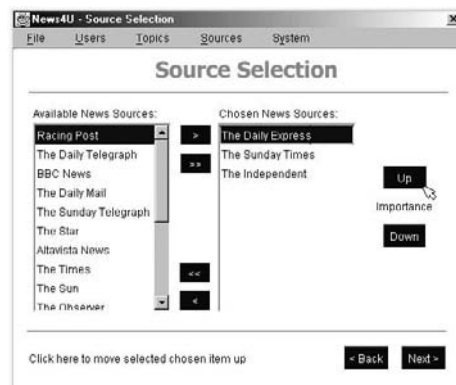


Fig. 1. User news source selection window.



Fig. 2. User topic selection window.

Personalised news delivery is a user-centered approach to news dissemination. As such the style and content of the news delivered should be controlled as far as possible by the user. Thus in our systems we allow the user to select from a wide variety of possible news sources such as national and international newspapers and internet news sources. Users can add new topics of interest either for permanent inclusion or for an individual session. Figures 1 and 2 show examples of source and topic selection windows from one of our systems.

After the preferred sources and topics have been chosen, a key feature of these systems is how to select the most relevant story on an individual topic. As a starting point to addressing this issue we are exploring the use of information retrieval techniques.

3.1 Information Retrieval

Each topic of potential interest to the user is represented by a profile. This is a set of search terms (conflated words) which describe the user's interests. A best-match retrieval matching score is calculated between a profile and each possibly relevant news story.

Preprocessing The news articles are originally contained within HTML pages which have a format specific to the news provider. The news delivery system first extracts the headlines and other features such as author, along with the story itself and formats this into a standardised local HTML page.

In order to improve the efficiency and effectiveness of the retrieval process each document is preprocessed. In the first stage each story document is filtered to remove stop words, such as common prepositions, which are not effective for retrieval. Next, words are conflated using the standard Porter suffix stripping algorithm [1] which standardizes words forms to form search terms which encourage matching between profiles and documents. Finally, the document terms are placed in an inverted file structure for efficient computation of the retrieval matching score.

Matching Score Calculation The simplest form of matching score for each query is simply to sum the common search terms between the profile and the document. Retrieval performance can be improved by applying term weighting to the document terms to give higher weights to items with good selectivity. Our systems adopt the Okapi BM25 probabilistic retrieval term weighting scheme [2] in which individual search terms are weighted using the following equation,

$$cw(i, j) = \frac{cfw(i) \times tf(i, j) \times (K1 + 1)}{K1 \times ((1 - b) + (b \times ndl(j))) + tf(i, j)}$$

where $cw(i, j)$ represents the weight of term i in document j , $cfw(i)$ is the standard collection weight (often referred to as inverse document frequency weight), $tf(i, j)$ is frequency of term i in document j , and $ndl(j) = dl(j) / (\text{Av } dl \text{ for all docs})$ is the normalised length of document j . $K1$ and b are empirically selected tuning constants for a particular set of documents.

The output of the matching process is a ranked list of news stories. The highest ranking document is included in the main personalised newspaper, but links are included to all stories on the same topic. Hence the user is given immediate access to the highest scoring story, but can easily pursue other articles on the same topic. Although not implemented in our current system we intend to extend this system to incorporate relevance feedback techniques to adapt user profiles and search term weights to better reflect user interests.

3.2 User Interface

Users are very familiar with the format of standard newspapers. For example, they expect to find certain section types distributed within the paper. Also newspaper pages are typically laid out to present the material in easy to read columns. Our efforts to replicate the practical features of newspapers, in particular to automate the layout of columns in the composition of a personalised paper, revealed some of the shortcomings of existing standard HTML markup. In order to realise the full possibilities of personalised news delivery making use of existing well defined layout concepts will require more sophisticated markup technologies perhaps relying on XML with appropriate browser software.

3.3 Evaluation

In small scale trials at our university we have generally received positive user feedback on our experimental systems. Users in general find the systems easy to use, although some users experienced difficulties understanding some of the concepts in the interface suggesting that we need to conduct more user studies in this area to make better use of concepts with which users are already familiar. One interesting observation is that once given a personalised news service users started to request access to more specialised news sources suggesting that they have quickly adapted to user-centred news delivery.

4 Concluding Remarks and Further Enhancements

There are many issues which would need to be addressed before commercial systems could be considered. For example, there are considerable issues of copyright and pricing of the stories, and at a practical level the relationship between information servers and local clients would have to be clearly defined.

In terms of technical enhancements there are two particular areas of interest. In the current systems individual articles are retrieved using the user's topic profiles. A desirable feature would be to automatically link together potentially related stories from different news sources on a particular day and also over time as a story develops. Techniques to perform story linking are being explored based on methods introduced in [3]. Also recent research has shown that multimedia news material can be retrieved effectively. Further work on personalised news delivery will seek to integrate radio and television news material.

References

- [1] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, July 1980.
- [2] K. Spärck Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and status. Technical Report 446, Cambridge University Computer Laboratory, August 1998.
- [3] N. Uramoto and K. Takeda. A method for relating multiple newspaper articles by using graphs, and its application to webcasting. In *Proceedings of COLING 98*, pages 1307–1313, Montreal, 1998.

Building a Digital Library of Web News

Nuno Maria and Mário J. Silva

Faculdade de Ciências da Universidade de Lisboa, Portugal
Departamento de Informática
{nmsm, mjs}@di.fc.ul.pt

Abstract. We introduce a new information system for organization of a Digital Library of news articles found on the Web, with automatic topic classification. We present our strategies to deal with different update frequencies of news Web sites, the classification methodology, the data model for storing news articles, measurements on the data retrieved and finally results of classification of this type of information.

Overview

The number of publications and news articles published on the Web had a dramatic increase over the last years. Readers often want to search and retrieve these publications for past news articles related to a particular subject or story. However, a Web news archive is usually not available and we can not rely in the publication archive. Search engines fail in coverage and often present broken links to publications. In addition, electronic publications separate their articles into a set of categories, but classifications are not uniform, leading to a poor satisfaction of readers information needs. Our project involves the creation of a framework to define Web services that let users see past published news on the Internet sites organized in a common category scheme.

To achieve this, specialized information retrieval and text classification tools are necessary. The Web news corpus suffers from specific constraints, such as a fast update frequency or a transitory nature, as news information is “ephemeral.” As a result, traditional IR systems are not optimized to deal with such constraints. As each publication has its own scheme of topics, it is also difficult to match the classification topics defined by each publication.

Our framework for Web news retrieval is built on broadly available research work [1]. In the automatic text categorization field, detailed examinations on the behavior of statistical learning methods and performance comparisons are periodically available [2, 7]. These studies show Support Vector Machines as the most efficient technique currently available. Recent work on Topic Detection and Tracking (TDT) and document clustering is also available [8]. These fields are studying automatic techniques for detecting novel events from streams of news stories and track events of interest over time. However, these areas are still a recent research activity and many research questions remain open.

In our work, we are addressing some of these problems applying advanced information retrieval and classification techniques to the physical world of Web

publishing, as we index and classify the major Portuguese news publications available on the Web.

In our research, we have identified the following main problems associated with retrieval of theme-based news:

- In general, news articles are available on the publisher's site only for a short period of time. Many Publications do not give access to their archive of previous editions and a database of references becomes easily invalid.
- Many news Web sites are built dynamically, often showing different information content over time in the same URL. This invalidates any strategy for incremental gathering of news from these Web sites based on their address.
- Direct application of common statistical learning methods to automatic text classification raises the problem of non-exclusive classification of news articles. Each article may be classified correctly into several categories, reflecting its heterogeneous nature. However, traditional classifiers are trained with a set of positive and negative examples and typically produce a binary value ignoring the underlying relations between the article and multiple categories;
- News clustering, which would provide easy access to articles from different publications about the same story, can be an important improvement. The automatic grouping of articles into the same topic requires very high confidence, as mistakes would be too obvious to readers.

To address the above presented problems we believe that it is necessary to integrate in a global architecture a specialized retrieval mechanism and a multiple category classification framework, including a data model for information and classification confidence thresholds.

We have been developing and evaluating a prototype of a system that addresses the above introduced requirements.

Architecture

Our system integrates a set of built or customized components for retrieval and classification of text documents. Figure 1 presents our system architecture for retrieval and classification of Web news.

The Retrieval Framework is implemented as a modified version of the Harvest System [1]. To deal with multiple publication periodicities, several queues are created and managed, expressing different update priorities and different gathering schedules. Information mediators [6] specialized for each publication are configured with rules to determine the URLs of the current edition and perform the collection of news articles according to the different schedules of each scanned publication. On a higher abstraction level, we view the Retrieval Framework as a service provider of a continuous stream of news articles retrieved from news Web sites. Each retrieved article is then delivered to our Classification Framework.

As articles arrive from the stream provided by the Retrieval Framework, they are automatically converted into a vector format according to a pre-defined vocabulary.

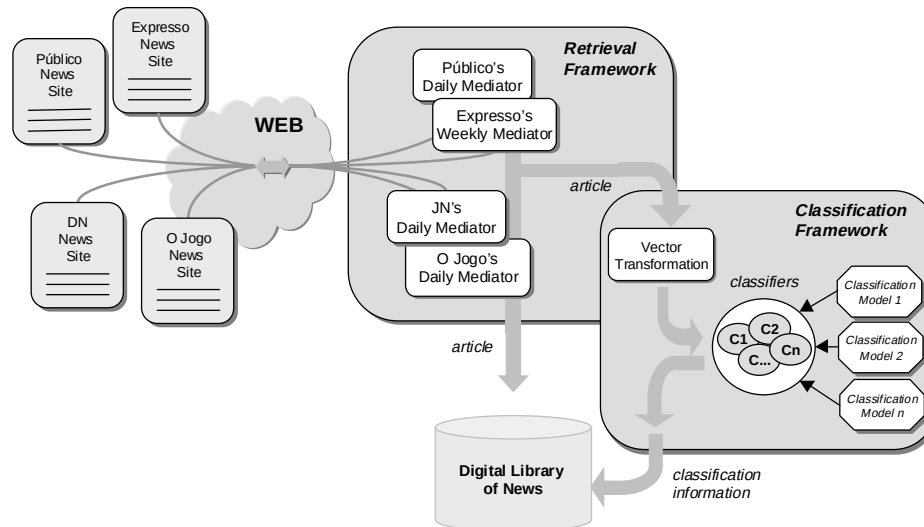


Fig. 1. Architecture for retrieval and classification of news articles. The Retrieval Framework collects articles from news sites distributed over the Web by specialized information mediators. These articles are sent to the Classification Framework for classification on a common scheme according to train examples provided by a local Digital Library. The generated classification information is then stored in our Digital Library of News for use by publications

The actual classification, based on pre-computed classification models, one for each pre-defined category, is performed with this vector format. The models are built with trained examples prepared from the news corpus provided by the local news library [4]. In our implementation this corpus is actually a daily newspaper with manually classified articles. The Classification Framework uses SVM^{light} , a package developed for automatic text classification using Support Vector Machines [3].

Once one article is processed, its vector format and classification are loaded in our Digital Library of News. It also stores the article's confidence level, returned from the classifier for each model, and possibly related articles. Proximity between articles is detected by a Similarity Detection mechanism. Each new article is compared to recently processed articles to check for similarities. If similarity is detected, then the article is grouped with other articles on a related cluster.

Results

The Portuguese online news corpus has 15 national news wires. We validated our system with a sample of 5 publications. Its overall performance on this sample is summarized in Table 1. We estimate in 4 GB annual storage requirements to maintain the text of the 5 publications. We believe these are acceptable storage needs considering today's hardware cost. Our strategies, restricting mediators to index only online "interesting" content, provide scalability to our solution making it a useful tool to small communities with specific interests.

Table 1. Measurements made in our Digital Library of News with 5 publications

Average Daily insertions	Average text+metadata size	Average text size	Average classification time
762 articles	12.0 KB / article	4.4 KB / article	1.0 sec / article

Our Classification Framework was validated with a set of articles from the publications sample representing a day of publishing activity (approximately one thousand articles). Here, we only outline the main results. A more detailed discussion is available in [5].

The classification mechanisms achieved 94,5% of accuracy with non-exclusive classification. Approximately 37% of the articles were classified in more than one of the eleven defined categories.

In the extremely dynamic environment of Web news we also must be aware of the degradation of the classifiers' accuracy in time. We observed that classifiers should be trained with a set of articles from a wide temporal range to minimize the effects of season events. In our work, we also developed a framework for detecting deficient classifiers and minimize accuracy degradation.

We are now applying and measuring the accuracy of clustering techniques to our system.

References

1. Bowman, C., Danzig, P., Hardy, D., Manber, U. and Schwartz, M.: The Harvest Information Discovery and Access System. *Proceedings of the Second International WWW Conference*. pp.763-771, 1994.
2. Dumais, S., Platt, J., Heckerman, D. and Sahami, M.: Inductive Learning Algorithms and Representations for Text Categorization. *Proceedings of the Seventh International Conference on Information and Knowledge Management*, 1998.
3. Joachims, T.: Making large-Scale SVM Learning Practical. *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.
4. Maria, N., Gaspar, P., Grilo, N., Ferreira, A. and Silva M. J.: ARIADNE - Digital Library Architecture. *Proceedings of the 2nd European Conference on digital Libraries (ECDL'98)*, pages 667-668, 1998.
5. Maria, N. and Silva, M. J.: Theme-based Retrieval of Web News. *Proceedings of the Third International Workshop on the Web and Databases (WebDB'2000)*. To be published as Springer LNCS.
6. Wiederhold G.: Mediators in the Architecture of Future Information Systems. *IEEE Computer*, pages 38-49, March 1992.
7. Yang, Y. and Liu X.. A re-examination of text categorization methods. *Proceedings of the 22th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, pages 42-49, 1999.
8. Yang, Y., Carbonell, J., Brown, R., Pierce, T., Archibald B. T. and Liu X.. Learning approaches for Detecting and Tracking News Events. *IEEE Intelligent Systems: Special Issue on Applications of Intelligent Information Retrieval*, 14(4), July/August 1999.

Automatically Detecting and Organizing Documents into Topic Hierarchies: A Neural Network Based Approach to Bookshelf Creation and Arrangement

Andreas Rauber, Michael Dittenbach, and Dieter Merkl

Department of Software Technology, Vienna University of Technology
Favoritenstr. 9 - 11 / 188, A-1040 Wien, Austria
{andi, mbach, dieter}@ifs.tuwien.ac.at

Abstract. With the increasing amount of information available in electronic document collections, methods for organizing these collections to allow topic-oriented browsing and orientation gain importance. The SOMLib Digital Library System provides such an organization based on the self-organizing map, a popular neural network model. In this paper, we present the GHSOM, which, based on the same concepts, allows an automatic hierarchical decomposition and organization of documents, which very intuitively reflects the organization typically found in (manually organized) conventional libraries. We present a case study based on a 3-month article collection from an Austrian daily newspaper.

1 Introduction

With the increasing amount of textual information available in digital libraries, methods for accessing and providing orientation in these vast information repositories have gained significant importance. While research in information retrieval provides us with sophisticated methods for selecting relevant subsets of collections, only few systems address the need for interactive navigation in topically organized libraries. Still, this way of organization is the most common and thus most natural one to users of conventional libraries.

With the SOMLib Digital Library [5] we developed a system focusing on this aspect of information access by automatically organizing documents by topic on a two-dimensional map using a popular neural network model, namely the self-organizing map (SOM) [3]. Documents covering similar topics are found in neighboring regions on the resulting SOM, which very much resembles the conventional bookshelf metaphor of document organization. While this approach allows easy orientation in an unknown text collection and facilitates the integration of distributed repositories [6], the resulting map-based metaphor poses some constraints on the size of the document collection that can feasibly be displayed on a single two-dimensional map. The more topics are present, the larger the according map has to be. Furthermore, the size of the map has to be determined in advance, which turns out to be impractical for unknown document collections.

What we would like to have is a hierarchical representation of the various topics in a document repository similar to conventional library organization of several floors, sections and shelves. Approaches like the *Hierarchical Feature Map* [4] provide such a hierarchical organization. Albeit, the architecture of the resulting system has to be determined in advance, i.e. it has to be specified how many layers we would like to obtain, as well as the size of the maps on each layer. This turns out to be difficult for document collections, where the precise nature and topic distribution of a collection is unknown. Furthermore, some topics are present more prominently than others, thus requiring more map space for representation. Various approaches address the problem of having to define the size of a SOM in advance, such as the *Growing Grid* [2], where new units are added to map areas where the data, i.e. the topics, are not represented at a satisfying degree of granularity. However, since all of these methods rely on the creation of a single flat map, they tend to produce huge maps limiting their usability for the display of and navigation in document libraries.

To overcome these limitations we developed the Growing Hierarchical SOM (GHSOM), which automatically creates a hierarchical organization of a set of documents. This allows the network architecture to determine the topical structure of the given document repository during the training process, creating a hierarchy of self-organizing maps, each of which provides a topologically sorted representation of a topical subset. Starting from a rather small high-level SOM, which provides a coarse overview of the various topics present in the collection, subsequent layers are added where necessary to display a finer subdivision of topics. Each map in turn grows in size until it represents its topic to a sufficient degree of granularity. Since usually not all topics are present equally strong in a collection, this leads to an unbalanced hierarchy, assigning more ‘map-space’ to topics that are more prominent in a given collection. This allows the user to approach and intuitively browse a document collection in a way similar to conventional libraries.

2 Creating Topic Hierarchies of Shelves

Text documents can be thought of as forming topical clusters in the high-dimensional feature space spanned by the individual words in the documents. The SOM is an unsupervised neural network model that provides a topology-preserving mapping from such a high-dimensional input space onto a 2-dimensional output space. A trained SOM thus represents a topical ordering of the documents, i.e. documents on similar topics are located close to each other on the map. This is comparable to what one can expect from a conventional library, where we also find books on similar topics located next to each other. Thus, the SOM offers by its very architecture an ideal way for the organization of document repositories.

Contrary to standard SOM training, the GHSOM starts by initially training a small 2×2 SOM. With only 4 units to represent the wealth of topics in a collection, the so-called *quantization error of a map (QE)*, a measure for the quality of a SOM, is rather high. Thus, additional units are added to the SOM to allow a finer representation of the various topics, until they are represented at

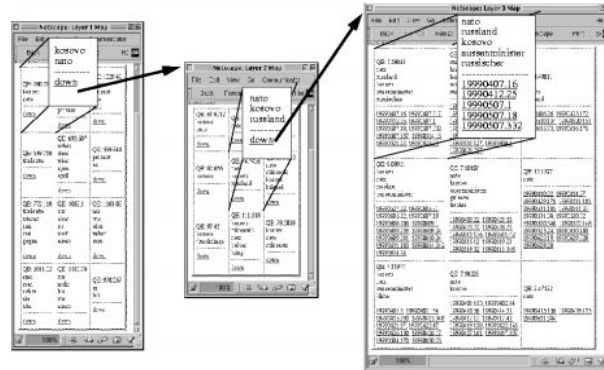


Fig. 1. Topic hierarchy on the Balkan War

a desired level of granularity. This is indicated by a threshold parameter τ_1 as a fraction of the initial QE . Next, a separate SOM is added for each unit of this first level SOM, which is used to represent the documents mapped onto the respective unit of the higher level map in more detail. Again, this map grows until it has a QE smaller than the fraction τ_1 of the *unit's quantization error* (qe) it derives from. The training process continues this way until the absolute granularity of data representation provided by a unit's qe exceeds a given threshold τ_2 , provided again as a fraction of the initial QE . (A more detailed analysis of the architecture and training process of this network model can be found in [1].) We thus end up with a hierarchy of individual maps, each representing a specific topic at a given level of granularity, allowing straight-forward browsing of and orientation in a document repository.

3 Walking through a Newspaper Archive

In the experiments presented hereafter we use an article collection of the Austrian daily newspaper *Der Standard* covering the second quarter of 1999. The 11.627 articles are represented by automatically extracted 1104-dimensional feature vectors of word histograms weighted by a $tf \times idf$ weighting scheme.

Since it is impossible to present the complete topic hierarchy of three months of news articles, we will concentrate on a few sample topical sections. The top-level map of the GHSOM, which evolved to the size of 3×4 units during training by adding two rows and one column, is depicted on the left in Figure 1. Here, we find an overview of the main subject areas such as the war on the Balkan on unit (1/1) (we refer to a unit located in row x and column y as unit (x/y) , starting with (1/1) in the upper left corner), automatically labelled with *Kosovo* and *NATO*. Austrian internal politics (1/2), economy (3/1), and the European Union (2/3) are other examples of main subjects. Cultural topics are located on three units in the lower left corner, representing sports, fashion and theater, respectively.

If we are interested in the Balkan War subject, we discover a more detailed map in the next layer (Fig. 1 middle), where every unit represents a specific sub-

topic of this subject. The top-row sections concentrate on the general political and social situation, while unit (3/1) deals with the situation of refugees. Slobodan Milosevic is the main topic of the sections mapped onto the units (2/3), (3/2) and (3/3) in the bottom right corner of the map, whereas unit (2/2) covers political negotiations to resolve the crisis and Russia's role in the Balkan War.

One layer down the hierarchy we reach the actual 'book-shelves', which, for the previously mentioned section on negotiations, represents in the top left corner unit (1/1) articles about Russia's inclusion into the so-called G-8 meetings on a safe corridor for expelled people, followed by reports on the consequences of the G-8 resolutions and comments by the German Foreign Minister Fischer on unit (2/2).

Similar topic organization can be found in the other areas of the GHSOM hierarchy with, e.g., the section on the European Union being composed of subjects such as the relationship between the EU and the USA on imports of US beef or penal taxes, or Austria-specific EU topics such as transit regulations or the agriculture subsidies within the Agenda 2000 pact.

4 Conclusions

The SOMLib Digital Library System has shown to be a very intuitive system by providing a topical organization of documents similar to conventional libraries and including a representation based on metaphor graphics. With the GHSOM, a dynamically growing hierarchical extension to the basic SOM model, we can now provide a hierarchical organization of topics, where each topic is assigned the required map space according to its prominence in a document collection. The resulting organization allows for an even more intuitive navigation and orientation within an unknown document collection. It allows users to find out, which topics are present, and where to find specific documents, putting the emphasis on browsing rather than mere retrieval, while retrieval functionality obviously can be integrated the same way as for the standard SOMLib system.

References

1. M. Dittenbach, D. Merkl, and A. Rauber. Using growing hierarchical self-organizing maps for document classification. In *Proc. European Symp. on Artificial Neural Networks (ESANN00)*, Bruges, Belgium, 2000.
2. B. Fritzke. Growing grid – a self-organizing network with constant neighborhood range and adaption strength. *Neural Processing Letters*, 2, No. 5:1 – 5, 1995.
3. T. Kohonen. *Self-Organizing Maps*. Springer Verlag, Berlin, Germany, 1995.
4. R. Miikkulainen. Script recognition with hierarchical feature maps. *Connection Science*, 2:83 – 101, 1990.
5. A. Rauber and D. Merkl. The SOMLib Digital Library System. In *Proc. Europ. Conf. on Research and Advanced Technology for Digital Libraries (ECDL99)*, Paris, France, 1999. LNCS, Springer Verlag.
6. A. Rauber and D. Merkl. Providing topically sorted access to subsequently released newspaper editions or: How to build your private digital library. In *Proc. 11th Int'l Conf. on Database and Expert Systems Applications (DEXA00)*, Greenwich, UK, 2000.

Daffodil: Distributed Agents for User-Friendly Access of Digital Libraries

Norbert Gövert, Norbert Fuhr, and Claus-Peter Klas

University of Dortmund {goevert, fuhr, klas}@ls6.cs.uni-dortmund.de

1 Introduction

The Internet makes searching for literature in Digital Libraries (DLs) feasible. However, often a user has to contact several DLs to satisfy a given information need. This leads to usability problems due to the heterogeneity of the DLs. One aspect is that the information structures of the systems differ. In fact, relevant information may be spread across several DLs. The other aspect of heterogeneity is differing browsing and searching functionality, of course presented to the user through different user interfaces and query languages.

We demonstrate *Daffodil*, a system of distributed agents for user-friendly access to Digital Libraries, which overcomes these problems. Daffodil integrates several DLs freely accessible through the Internet to present to the user a single virtual DL. In addition to unifying access points and user interfaces Daffodil also produces synergies while exploiting different DLs: information from different DLs about the same document in different DLs is merged. Further Daffodil combines the functions of the underlying DLs in order to provide *high-level* search and browsing functions. As a systematization of Daffodil's functionality we use Bates characterization of information search activities on four levels of abstraction [2]:

Moves are understood as basic units in the information search process. They correspond to the commands of information systems. Examples are adding a term or a condition to a given query or following a link.

Tactics comprise one or more moves with the purpose of improving or speeding a search. One tactic is query refinement, e. g. generalizing or specializing a search term to improve recall or precision, respectively.

Stratagems exploit information structures of a given search domain, e. g. in a journal run (browsing a journal which is central to a given information need), in a citation search (following links in a relevant document), or in an author search (searching for other publications of an author, who occurred earlier in one or more relevant documents).

Strategies are plans which comprise moves, tactics, and stratagems, in order to process a complete information search, i. e. cover an information need. A simple strategy for literature search in a novel area would be an area scan in a first step, followed by an author search with authors of relevant documents found in the first step.

Most DLs provide functions on the level of moves only. From that level we can distinguish the other levels by their orientation towards a predefined target (i. e. solving a specific information search task) as well as by their *strategic* element: Tactics, stratagems, and strategies describe *plans* by which their respective target can be reached. Daffodil offers search and browsing functions on a strategic level in order to guide users

through the information structures. The following section describes the architecture of Daffodil, which implements such search activities. In section 3 we give a brief sketch of Daffodil's user interface.

2 Layered Architecture of Daffodil

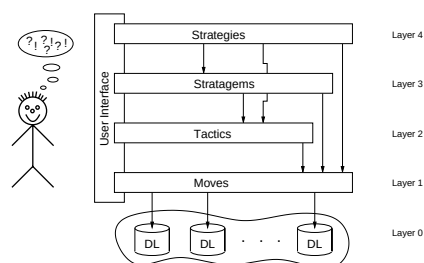


Fig. 1. Architecture of Daffodil

Figure 1 illustrates the layered architecture of our system. The bottom layer is made up of the digital libraries connected to Daffodil, each of them with its own interface. The layers one to four are the *moves*, *tactics*, *stratagems*, and *strategies* layers, each providing the respective class of functions to the user via the user interface.

For implementing the layered architecture, we developed a system of interacting distributed agents. The flexibility and extensibility gained from using an agent-based architecture enables us for example to integrate new DLs and new search services. Agent characteristics like *ability to communicate*, *adaptiveness*, and *autonomy* [4] allow for the design of such a system.

Moves

The moves layer deals with heterogeneity on a syntactical level. It provides for a uniform interface to the DLs of layer 0. For this purpose a *wrapper agent* has been implemented for each DL connected to Daffodil. The main task of a wrapper is to map the functions and information structures of the associated digital library into the system of interacting agents. Twelve DLs have been wrapped into the system so far, each with its different emphasis. For example we use CiteSeer¹ for citation searches and as a full-text service, as well as the NCSTRL² service; the ACM DL³ can be used for journal and conference runs; any of the DLs connected to Daffodil can be used for performing the standard task in Information Retrieval, i. e. searching for documents by specifying one or more query conditions w. r. t. document content and associated attributes.

Tactics

The tactics layer deals with heterogeneity on a semantical level. It produces synergies in that it combines the different classes of information objects of the DLs. Following the ideas of *multi-level hypertext* [1], we can arrange them on four levels of abstraction [3]: On the *schema* level, different attributes of documents and their respective metadata are modeled. Examples for standard schemas are the *Dublin Core* or bibliographic schemes like e. g. MARC. In Daffodil we use a slightly extended version of the BIBTEX attribute set. The *attribute value* level comprises the domain of a given attribute, e. g. a list of author names. Often such attribute domains are structured themselves, e. g. in thesauruses or classification schemes. A metadata record on the *metadata* level subsumes the attribute values of a given document w. r. t. a given schema, e. g. bibliographic attributes

¹ <http://www.csindex.com/>

² <http://www.ncstrl.org/>

³ <http://www.acm.org/dl/>

like author, title, and publication year. Finally, the *full-text* level contains the documents themselves, usually in formatted form, such as PDF or PostScript.

Search and browsing functions offered by DLs can now be identified by means of intra-level links and links between different levels. For example, an *attribute link* connects an attribute value (e. g. an author name) with documents and/or metadata records to which the value is assigned; so users may first browse the list of attribute values (e. g. the author index) and then view publications with a selected value. An example for intra-level links are *similarity links*. They exist between similar objects, either attribute values (e. g. *see also*-terms from a thesaurus) or documents; a user may look for documents which are similar to a known relevant one.

DLs can now be characterized by the levels on which information objects are provided and by the links they offer on or between these levels. Daffodil's tactics layer extends the functions of the connected DLs in that it combines the information structures of several DLs and the respective (intra-level) links into one virtual Digital Library.

Stratagems

As described in the previous sections, DLs provide a large variety of basic search and browsing functions, depending on the types of information objects they provide. They require that users are supported in using these functions effectively.

With the layers one and two we get a uniform interface for accessing the functions and information objects of the DLs. Combining them in a single system, we now can offer additional and higher-level search and browsing functions, thus providing a user-friendly access to the information structures. The stratagem layer aims at guiding the users through information structures in a given search domain. On this layer we implemented agents for seven stratagems, under them *author search* (searches for publications by focusing on authorship; users can browse authors, search for co-authors and extract authors from a set of metadata records in order to formulate new queries), *area scan* (which supports navigation through a classification scheme and provides links to documents belonging to each category), and *aggregation browse* (which focuses on tables of content of journals and conferences).

Strategies

Since strategies are highly dependent from the situation in which a user performs an information search, it is difficult to provide for a fixed set of strategies. Therefore we spent little effort here and only provided the *presearch interview* strategy which is applicable in the most general situation. The user is asked for known facts about the information objects to be searched, e. g. whether specific authors, journals, or conferences are known which might be helpful for the information search. Depending on the answers, Daffodil proposes one or more stratagems to process the user's information search task.

3 The User Interface

The functions implemented in the layered architecture of Daffodil are offered to the users through the user interface. In our system the user interface is implemented by using HTML with forms and frames, thus it can be easily accessed via a Web browser. The task of the user interface is twofold: on the one hand it must present the overall functionality of the agent system to the user; on the other hand it has to visualize the results adduced by the system.

The interface is divided into four main frames:

- Upper Left:** Query condition builders. 'Active query conditions' shows 'Author' and 'Cluet'. 'Inactive query conditions' shows 'Author' and 'Milo'. Buttons: 'submit', 'reset', 'apply'.
- Upper Right:** 'Add new term' input with 'author' dropdown and 'add' button. A list of authors: Sophie Cluet, Serge Abiteboul, Tova Milo, Olga Kapitskaja, Divesh Srivastava, C. Delobel, G. Moerkotte. Buttons: 'Cosauthor search', 'Author search'.
- Lower Left:** List of document metadata records. Each record includes a checkbox, title, authors, and conference. Examples include 'Nested Queries in Object Bases differences between unnesting in the relational and unnesting in the' and 'A database interface for file update'.
- Lower Right:** 'Document Details' for a selected record. Fields include Title, Author(s), Conference, and Year. A section 'Actions on this document' lists actions like 'Complete this document metadata', 'Find cited documents', 'Find documents citing this document', 'Open document fulltext in a separate window', and 'Find similar documents'. A section 'Actions on picked values' lists 'Query other documents with picked values instantly' and 'Refine your query with picked values'. A 'Document is relevant?' section with 'Yes' and 'No' radio buttons.

Fig. 2. User interface visualizing an author search

list of metadata records. Selecting a record displays additional information about that document in the lower right frame. The upper right frame shows the domain of a single attribute, in this case a list of author names, and allows for filtering and browsing the attribute values. Each of the frames comprise a set of functions, which can be performed on the information objects displayed, i. e. these functions implement intra-level links and links between the different levels of the multi-level hypertext.

4 Summary

We present Daffodil, which implements a virtual Digital Library, comprising information structures and functions of a (extensible) set of Digital Libraries freely accessible through the Internet. Daffodil provides for user-friendly access to contents and functions of Digital Libraries in that it offers high-level search and browsing functions to the users. Such functions guide the user through the information structures in a single-minded manner.

References

- [1] M. Agosti, G. Gradenigo, and P.G. Marchetti. Architecture and functions for a conceptual interface to very large online bibliographic collections. In *Proceedings RIAO 91*, pages 2–24, Paris, France, 1991.
- [2] M. J. Bates. Where should the person stop and the information search interface start? *Information Processing and Management*, 26(5):575–591, 1990.
- [3] N. Fuhr. Information retrieval in digitalen bibliotheken. In *21. DGI-Online-Tagung – Aufbruch ins Wissensmanagement*, Frankfurt, 1999. DGI.
- [4] Hyacinth S. Nwana. Software agents: An overview. *The Knowledge Engineering Review*, 11(3), 1996. <http://www.cs.umbc.edu/agents/introduction/ao/>.

Figure 2 shows an author search. The screen is split into four main frames. The upper left frame presents the current query, consisting of a set of conditions. The other three main frames represent one level of the multi-level hypertext each. The lower left frame shows the query result, a

An Adaptive Systems Approach to the Implementation and Evaluation of Digital Library Recommendation Systems

Johan Bollen and Luis M. Rocha

Los Alamos National Laboratory
Computer Research and Applications Group
Los Alamos, NM (USA)

Abstract. The focus for information retrieval systems in digital libraries has shifted from passive repositories of information to recommendation systems that actively participate in retrieving useful information, and can furthermore learn from the retrieval behavior of users. We propose a novel evaluation methodology for such systems based on the concepts of shared knowledge structures, and system development reliability and validity.

1 Evaluation of Recommendation Systems

1.1 Precision and Recall for Traditional IR

Traditionally, IR evaluation methodologies have focused on the assessment of the set of documents retrieved or recommended to users [2]. Performance evaluation of IR systems, both in experimental [3] and analytical [4] approaches, is based on measures that use the sets of retrieved and relevant documents: $RETR$ and REL . The most common measures are $recall = \frac{RETR \cap REL}{REL}$ and $precision = \frac{RETR \cap REL}{RETR}$. The problem lies in determining the elements of the relevant set REL , which ultimately depends on the subjective judgments of human experts.

Traditional IR performance evaluations avoid the inherent subjectivity of relevance evaluations, even when techniques such as cross-validation, expert sampling (see e.g. the TREC (Text REtrieval Conference) standard [3]) and query expansion techniques [5] are used. They can therefore not take into account the requirements of specific communities of users.

1.2 Evaluation of Recommendation Systems

Interactive, adaptive RSs, contrary to traditional IR systems, actively recommend information items and adapt the relational information for the set of documents and keywords they operate on to their user communities' characteristics. For these kinds of systems we are mostly interested in the conditions that allow the associations among documents, keywords, and between the two, to reflect the knowledge of their particular community of users: *validity*, and if and how the adaptive algorithms of RS converge to the knowledge of users: *reliability*. We therefore propose a simulation approach that enables us to assess how recommendation systems interactively generate representations of the knowledge of their community of users, rather than attempting to describe the semantics or relevance of individual documents, as current IR methodologies do.

1.3 Validity and Reliability with Shared Knowledge Structures

Collective choice and shared knowledge structures We have taken an approach in which we operationalize the aggregation of user knowledge as shared knowledge structures (SKS). From the comparison of the recommendation system's knowledge base and the SKS, we can then devise stability and validity performance measures of the recommendation system's adaptive behavior. The work of Richards [6] on collective choice has been particularly relevant in this context. Richards' model is based on the assumption that a population of agents needs to make a collective choice on a set $\mathcal{A}$ of alternatives a_i on which a shared knowledge structure (SKS) is imposed. The SKS will be represented by a non-directed, connected graph $W(A)$. This graph is understood as the particular associative knowledge structure that all agents share in an environment. Agents are assumed to prefer one of the alternatives a_i , thus, for each agent, there is a preference partial order $D_i(a)$. Rather than modeling the capacity of a collection of agents to reach a single choice, we are interested in modeling an adaptive recommendation system's ability to reliably adapt to a valid representation of the patterns of associations in the agents' SKS.

A simulation methodology We propose the following simulation approach to a performance evaluation of an adaptive recommendation system. Let K and D be, respectively, a set of keywords and documents that are derived from a number of given database records and combined into a large set of alternatives $\mathcal{A}$. A particular community of users utilizes a subset of $A \subset \mathcal{A}$. The SKS of this community of users then is $W(A)$. To model such a community, a population of agents is generated with partial orders $D_i(A)$ starting on a particular element $a_i \in A$. Because A is typically very large, we use incomplete partial orders D_{ip} , that is a partial order starting at a_i , but extending only to the nearest levels p . Parameter p indicates how much of the SKS a given agent "knows" about.

This community of agents is then set to interact with a given recommendation system which will aggregate the partial knowledge of all agents into the adaptive structure of the information resource. We start with an initial graph $V(A)$ representing the initial associative structure of the information resource we wish to model.¹ We then let the population of agents that share a SKS $W(A)$ interact with $V(A)$, using any number of adaptation algorithms. In this simulation methodology, validity can be defined as the similarity between the final structure of $V(A)$ (after completion of a sufficient number of iterations of the adaptation algorithm) and $W(A)$. The reliability measure requires a number of in parallel developed $V(A)$ s, to determine whether the adaptive algorithm converges on one and the same associative structure in different runs, given the same population of agents.

2 The Identification of User Communities

The above described methodology relies critically on our ability to generate a population of agents that share a SKS. As a demonstration of the viability of the proposed methodology, SKSs for a number of communities the Los Alamos National Laboratory's (LANL) Research Library (<http://lib-www.lanl.gov/>). have been generated based on keyword and document semantic proximities [7].

¹ This can be an initially random structure for information resources that have no initial set of document links such as citations, etc.

2.1 User Log Clusters

The Research Library at the LANL is a strongly networked digital library, i.e a large part of its repository has been moved to a digital format and can be downloaded from the library's web site. Extensive web logs keep track of users' retrieval patterns. A row-normalized similarity matrix for journal titles had previously been generated based on the co-occurrence of journal titles in user retrieval paths in the February 1999 Research Library web logs (for a more detailed description of this technique, see [1]). A hierarchical cluster analysis was performed on this matrix, revealing a number of persistent journal clusters. Three clusters were selected for further analysis based on their size, subject and stability over different thresholds (see fig. 1): *Information Sciences* (cluster1), *Molecular Biology* (cluster2) and *Non-Linear Science* (cluster3).

cluster 1	cluster 2	cluster3
Inform. Sci.	Mol. Biol	Non-Lin. Sci
Inform. Proc. Letters	Analyt. Chim	Prev. Med.
Inform. Sciences	Tetrahedr. A	Chaos
	J. Mol. Biol.	Siam J. Comp
	Tetrahedron	Quant. Res.
	Tetrahedron L.	Siam Rev.
		Arch. Env. C.

Table 1. Three selected LANL RL user-based journal clusters derived from the library's web log co-occurrences

A user community was derived by determining a set of users that had frequently downloaded articles published in the journals in the journal clusters. For each of those communities a list of the 20 most frequently articles and their associated keywords was compiled ².

As defined in [7], Semantic Proximity values were calculated for all pairs of articles and associated keywords, resulting in a SKS graph for each user cluster. Fig. 1 shows a graph that represents the keywords relational information for the community of users in cluster 3, i.e a measurement of this group's SKS.

After having derived a set of different SKS like the one shown in fig. 1, we can derive a set of general properties for SKSs (e.g. metricity, transitive closure, etc.) and consequently automate their derivation from digital library web resources. From this preliminary data, we plan to generate a set of agents according to the methodology described in section 1.3, and used these for a first series of simulations.

3 Conclusion

We presented a methodology that based on the concept of shared knowledge structures will enable us to measure and describe the performance of active recommendation systems.

Future research will focus on applying this methodology to a number of existing active recommendation systems that can lead to the following applications. First, this evaluation methodology can be used to gather empirical data on the knowledge of communities of researchers within an institution. Second, measurements of the speed of

² this process will be further detailed at the ECDL2000 conference presentation

Are End-Users Satisfied by Using Digital Libraries?

Mounir A. Khalil

The City College of New York,
New York City, United States of America
Moucc@cunyvm.cuny.edu

Abstract. There are many books and journal articles written and published about the Digital Library (also denoted as Electronic library or Virtual Library which present the definition, description, components, usefulness, etc. of the Digital Library – but nothing has been mentioned or written about the satisfaction of the users for accessing needed information. A questionnaire was developed to survey the behavior of end-users and measure their understanding levels of the meaning of the Digital Library.. Below are the results of the global survey. Electronic copyright and licensing as well as their effects upon research and education are discussed.

1 Introduction

The topic of digital libraries is of great interest to computer network engineers, systems engineers, computer scientists as well as librarians in recent years. There are many researches and initiatives in the United States and Europe and in other parts of the world including Asia, Africa, and Australia. The National Science Foundation announced about an international program on digital libraries. Librarians became interested in knowing how useful is the digital library to their clientele. So I developed the surf to collect data from library users in order to answer the following questions:

- How can end-users define and understand digital libraries?
- How useful are digital libraries to educators, students, and researchers?
- Are there any advantages in using the digital libraries over the traditional libraries?
- What disadvantages do digital libraries have?
- Will education, research and teaching improve when digital libraries are in full place?

The questionnaire was posted on many Listservs (ASIS-L, Lita-L, DDS-L, PSC-L, BUSLIB-L, CIRCPLUS-L, DIGILIB-L, LIBADMIN-L, LITA-L, IFLA-L, SERIALST-L, CULIBS-L, METRO-L, IFLA-L, ...etc as well as the Internet. Quick responses came from different end-users from different countries such as United Kingdom, France, Netherlands, Italy, Austria, Spain, Poland, Brazil, India, Pakistan, China, Hong Kong, Australian, and Unites States of America.

2 Findings

The collected data from the survey will cover the following items:

Types of libraries and librarians,
Definitions of DL as perceived by end-users,,
Its components, methods of accessing electronic information,
Advantages and disadvantages of DL
Improving education .
Electronic Copyright Laws.

2.1 Types of Libraries

- Academic 58%
- Business 10%
- Special 8%
- Medical 5%
- Law 3%
- School 8%
- Other 8%

It is obvious that academic libraries (college and university libraries) are using DL followed by business, special, medical libraries.

2.2 Definitions of DL from the Survey

The respondents gave many definitions of the Digital Library. We narrowed down to the common agreement by many respondents:

1. A collection of digitized material available electronically especially but not exclusively over the Internet.
2. A library without walls with electronic access to documents whether originally published in paper or electronically—which were available in the traditional one.
3. A Digital Library is the virtual library which users can get access to information by transcending the ordinary barriers of time and space.
4. A library where all information and data can be accessed from remote sites without placing a foot in other locations.
5. A library or storage of information resources in digital format such as CD-ROM, PC or other networked devices, or the Internet or local server.
6. A library which the modern equivalent of the physical library.
7. A library which contains digitized resources from valuable, special collections , digital images , audiovisuals, and traditional print that has been digitized
8. A place on the web (WWW) which provides full image to a variety of formats — video, sound, text

2.3 Digital Library Components

The respondents indicated that main components of the Digital Library are:

- Powerful workstations or servers	75%
- Advanced software	66%
- Electronic publishing	58%
- Others (print materials, scanners, sound, graphics)	17%

2.4 Methods of Accessing Information

Accessing electronic information through the Internet was the first choice and others as follows:

- The Internet	93 %
- Online databases	78%
- Business accounts	15 %
- Shared access usage	13 %
- Downloading :	
- Computers	70%
- E-mail	66%
- Printers	63%
- Diskettes	45%
- FTP protocol	8%

Some respondents mentioned that the cultural and technological factors are barriers in accessing information from other DL in other foreign countries.

2.5 Advantages

Respondents stated the advantages of using the Digital Libraries

1. Access from any location via Internet Service Providers (ISP).
2. Fast downloading and speedy dissemination to many users.
3. Save space and time.
4. No limited hours for usage -(available 24 hrs a day 7 days a week).
5. Current and updated information available.
6. Copies as PDF files are better than photocopies.
7. E-Journal articles on demand.
8. Preservation and security of original documents and files.

2.5 Disadvantages and Barriers

Downtime which is the biggest complaint.

Telecommunications costs

Need trained technician

Different hardware and software

Lack of electronic media

No universal standards for research and retrieval
Licensing restrictions and fees

2.6 Improving Education

75 % of the DL users confirmed that education would improve when digital are completed and widely used. About 5% of respondents had negative response and 25% gave no answers. Among the encouraging statements are:

DL will enhance teaching and research in many disciplines because there are not enough resources in traditional libraries. Indeed, there many benefits to faculty, students, and researchers.

3 Impact of Electronic Copyright Laws

68% of respondents stated that electronic copyright laws affecting electronic library will limit their use. But 8% gave a negative answer and 25% remained without any response. Many end-users positively do not agree with any electronic copyright laws on the entire Internet. Some expressed their views that economic factors tend to promote innovations that will work towards avoiding any restrictions that may come. In addition, publishers would be afraid about losing print revenue would tend to support legislation wholeheartedly.

4 Conclusion

This paper has showed the interest of end-users in using digital libraries. This survey has helped us in understanding the capabilities and limitations of digital libraries. Nowadays libraries have witnessed many changes. It is very clear that users still perceive digital libraries as electronic databases or electronic storage of information. There are differences in vision between what librarians and information professionals and what scientific and engineering researchers hold the definition of digital libraries. In addition, there is an extreme interest in using digital libraries in countries other than in the Western Hemisphere. They perceive advancing technologies as powerful tool for information.

CAP7: Searching and Browsing in Distributed Document Collections*

Norbert Fuhr¹, Kai Großjohann¹, and Stefan Kokkelink²

¹ University of Dortmund {fuhr,grossjohann}@ls6.cs.uni-dortmund.de

² University of Osnabrück skokkeli@mathematik.uni-osnabrueck.de

Abstract. This paper describes CAP7, a system for searching and browsing in distributed document (metadata) collections. The system architecture is similar to Harvest, comprising gatherer components and a retrieval engine; but instead of the limited SOIF data format, we use RDF and XML. The gatherer creates RDF metadata descriptions of collected resources. Before delivering the data to the retrieval engine, the RDF is transformed into valid XML. The query language supported by the retrieval engine is an extension of XQL with weighting and data types with vague predicates. The user interface provides for browsing as well as simple searching.

1 Introduction

The Math-Net project [3] is building a global information and communication system for Mathematicians. The system consists of several databases, one of which is MPRESS¹, a preprint index comprising 40 000 metadata records, about half of which contain Dublin Core metadata [4] encoded within the HTML ‘meta’ element. Although easy to use, the ‘meta’ element is not suitable for representing complex metadata. For example, for documents coauthored by two persons with different affiliations, the metadata representation would not exclude false coordinations of names and addresses. Therefore, the Math-Net project now employs RDF [1] as the underlying metadata format.²

The Math-Net project extensively uses the Harvest³ system. SOIF, the basic data format for the Harvest system, consists of attribute/value pairs and is thus basically equivalent to the HTML ‘meta’ element. Therefore, a new data format is needed for the Harvest system, and the CAP7 system extends the Harvest architecture to accomodate RDF.

* This work is part of the CARMEN project which was funded by the Global-Info initiative of the German government. See <http://www.mathematik.uni-osnabrueck.de/projects/carmen/index.en.shtml>

¹ <http://MathNet.preprints.org/>

² <http://www.mathematik.uni-osnabrueck.de/cgi-bin/MMM3.0.cgi>

³ <http://www.tardis.ed.ac.uk/harvest/>

The Harvest retrieval engine only supports flat SOIF structures. This is not sufficient for retrieval in XML [2] documents. Furthermore, Information Retrieval research has shown that for effective searching, vagueness and uncertainty are core concepts which should be supported by the retrieval engine. (For example, a user might not know the exact spelling of a personal name, or the exact classification code for a document.) These two concepts also imply that support for different data types (with vague predicates) is needed, i.e., personal names and classification codes in this example.

DesIRE⁴, the retrieval engine deployed in CAP7, supports searching in XML documents, with different data types and vague predicates.

2 Gathering and Summarizing

The gathering component fetches documents from relevant Web sites and summarizes them, producing metadata records in the RDF format (Resource Description Framework). For retrieval, these records are exported as XML documents to the retrieval engines. Below, we describe the different tasks in more detail.

Import. The gatherer retrieves information using access methods like HTTP, FTP, NNTP, or local files. The gathering process is customizable, for example restricting it to specific hosts or limiting the depth of following hyper links. The resources thus collected are heterogeneous: they may have different content and quality, for instance home pages, technical reports, or dissertations. On a lower level, heterogeneity is caused by the different formats, such as postscript, PDF, HTML or RDF. The summarizer extracts relevant metadata and produces an RDF description which differs in quality due to the heterogeneity. The best case is a structured resource with embedded Dublin Core metadata, and the worst case is a resource without any structure, e.g. a postscript document.

Storage. The extracted metadata is stored in an RDF repository. Since the structure of RDF is basically a directed labelled graph, the RDF repository can be thought of as a semantic network. The information represented by this network grows with each RDF description that is added to the repository. The complete information available from the gatherer originates from the gluing of the RDF graphs.

One major benefit of RDF is the possibility to use different vocabularies for the description of metadata. Therefore new information (e.g. generated by an inference engine) can be added in a well-defined way to the repository without affecting the original metadata.

Export. The complexity of the information structure represented by the RDF graph in the repository makes a direct processing by a retrieval engine difficult and is not suitable for a user-friendly retrieval. Hence the gatherer must be able

⁴ Dortmund extensible structured Information Retrieval engine

to break up the graph structure and to create a simplified tree-like data structure before exporting the data to retrieval engines in XML format.

3 IR/HT Engine

DesIRE, the retrieval engine deployed in CAP7, supports searching in XML documents, with different data types and vague predicates. Since a digital library usually contains both metadata and fulltext, using XML is more appropriate than RDF; it supports both kinds of data. Also, it is more efficient to search in XML documents (which have tree structure) than in RDF documents (which are directed graphs), even if the XML documents do contain a few links.

All documents should have the same structure so that users know how to formulate their queries. (As a simple example, suppose that each document would use its own markup for the author name. This would basically require the user to formulate a query for each document, which is clearly impractical.) Requiring all documents to conform to the same DTD is an easy way to ensure that their structure is similar. Therefore, we require valid (rather than just well-formed) XML documents.

XQL [5] provides good support for searching the structure of XML documents, but the data type support is weak and there are no vague predicates. XIRQL, the query language used in DesIRE, augments XQL with various data types and vague predicates (weighting). Due to an extensible design, it is possible to add more data types and (vague) predicates for additional applications.

Description of XQL. An XQL query is applied to an XML document (which is considered a tree) and returns a set of nodes (hence called ‘target nodes’). A query specifies a path from the root node to the target node. For each node on that path, the query may specify the node type (element node, attribute node, text node), as well as the (element or attribute) name, and the element content (or attribute value). In addition to equality, comparison operators are provided. For full complexity, it is possible to also specify a condition which must be fulfilled by the subtree rooted at that node. (A condition is like an XQL query and is considered true if it returns at least one node.) Using a wildcard operator, it is possible to allow an arbitrary number of intervening nodes on the path.

Weighting. Whereas in XQL, the result of any query is a set of nodes (from the XML document tree), the result of a XIRQL query is always considered to be a weighted set; the weight attached to each node gives the probability of that node being relevant to the query. All XIRQL predicates are defined to produce such weighted sets as their output, and the combining operators (Boolean ‘and’ and ‘or’ and the path operators) accept weighted sets as inputs. XIRQL extends the Boolean operators with a weighted-sum operator; a query as a linear combination of terms has proven successful in Information Retrieval.

In addition to the binary comparison predicates (equality, less-than, greater-than), XIRQL provides vague predicates such as sounds-like (for person or place

names), similar-to (useful for many data types), and contains-term (for textual data types).

Data Types. The database administrator enriches the information available in the DTD with data type information and specifies the possible predicates and access paths. For example, all text in `author` elements could be declared a personal name, offering phonetic similarity search using the well-known soundex algorithm.

Result Presentation. The presentation is based on XML and XSL. The retrieval engine returns an XML document which represents the query result, which is then formatted for output using XSL. This method allows the database administrator great flexibility in designing the user interface for each application, by configuring the relevant XSL stylesheets.

4 Status and Outlook

Right now, six months into a two-year project, there is a prototype which provides searching in XML documents using a simple HTML form. The gatherer component is based on the Harvest-NG⁵ library. Currently, it only supports the processing of HTML documents using the XML representation of RDF for the handling of the generated metadata records. The next goal is to support more formats and to enhance the RDF features of the gatherer. For browsing, our system implements the concept of multi-level hypertext, supporting navigation in and between four different levels: schema, attribute values, document metadata, and fulltext. The retrieval engine already supports data types and vague predicates, but it is not yet possible to formulate queries with respect to the structure of the documents. Instead, the documents are transformed into a flat structure (where a document is considered to be a set of attributes) with XPath expressions, and retrieval is possible w.r.t. these attributes only. Full support of XIRQL is under development.

References

- [1] D. Brickley, E. Miller, B. Schloss, D. Singer, and R. Swick. Resource description framework (rdf) home page, August 1999. <http://www.w3.org/RDF/>.
- [2] Dan Connolly, editor. *XML: Principles, Tools, and Techniques*, volume 2 of *World Wide Web Journal*. O'Reilly, Sebastopol, California, 1997.
- [3] W. Dalitz and W. Sperber. The mathnet home page: Internet information services for mathematicians, January 2000. <http://www.math-net.org/>.
- [4] DCMI. Dublin core metadata initiative, home page, 2000. <http://purl.org/DC/>.
- [5] J. Robie, J. Lapp, and D. Schach. Xml query language (xql). In *QL'98 — The Query Languages Workshop*, December 1998. <http://www.w3.org/TandS/QL/QL98/pp/xql.html>.

⁵ <http://webharvest.sourceforge.net/ng/>

Representing Context-Dependent Information Using *Multidimensional XML*^{*}

Yannis Stavrakas^{1,2}, Manolis Gergatsoulis¹, and Theodoros Mitakos¹

¹ Institute of Informatics & Telecommunications,
National Centre for Scientific Research (N.C.S.R.) ‘Demokritos’,
153 10 Aghia Paraskevi Attikis, Greece.

² Knowledge & Database Systems Laboratory
National Technical University of Athens (NTUA), 157 73, Athens, Greece.
{ystavr,manolis}@iit.demokritos.gr
teo@softlab.ntua.gr

1 Motivation and Approach

XML (eXtensible Markup Language) is emerging as a new standard for data representation and exchange over the Web [3]. It is a markup language that resembles HTML, but unlike HTML, it focuses on the structure of data rather than on their presentation. The extensibility of XML makes it an ideal candidate for integration and manipulation of Web data through a common data model. A large number of DTDs (XML Document Type Definitions) that target all sorts of information domains has already been developed, and new DTDs are released at a fast pace. XML claims to be the enabling technology for application interoperability and for a unified view of heterogeneous information.

However, the nature of the Web poses a number of new problems. On the one hand, in traditional databases and information systems the number of users is more or less known and their background is to a great extent homogeneous. On the other hand, Web users do not share the same background and do not apply the same conventions when interpreting data. Such users can have different perspectives of the same entities, a situation that should be taken into account by Web data models. A similar problem appears when integrating information from various sources, where the same conceptual entity may exhibit different structure, or contain conflicting data.

Those problems call for a way to represent information entities that manifest different facets whose contents can vary in structure and value. Such cases are frequent in the frame of WWW where information providers cannot assume much about the needs and background of information consumers. As a simple example imagine a report that must be represented at various degrees of detail and in various languages. A solution would be to create a different XML document for every possible combination of variations. Such an approach is certainly not

^{*} This work has been partially supported by the Greek General Secretariat of Research and Technology under the project “Executable Intensional Languages and Intelligent Applications in Multimedia, Hypermedia and Virtual Reality” of *IIENEΔ*’99.

practical, since it involves excessive duplication of information. What is more, different variants are not associated as being parts of the same entities. Therefore, although XML is in principle capable to represent such multifacet entities, it falls short for a number of reasons, namely (a) hidden semantics: XML does not address directly the issue of multiple facets, therefore it is the responsibility of an application to assign semantics in an ad-hoc manner, (b) cumbersome syntax: representing multiple facets of an entity in XML cannot be done in an elegant way, and (c) duplication of information: information that is common is duplicated, which is undesirable.

In this paper we describe an extension of XML, called *Multidimensional XML* (MXML) [6] that incorporates elements of Modal Logics [4] in XML in order to tackle the aforementioned problems. In an MXML document each information piece may be qualified by a *context specifier* denoting the *world* in which the information piece holds. There is a number of alternative worlds for a document, generated by the *dimensions* used in that document. The idea of applying dimensions to languages is not new [2,5]. GLU [2] is a programming language based on the notion of multiple dimensions and ISE [7] is a multidimensional version of Perl. Our work was influenced by *Intensional HTML* (IHTML) [8], a Web authoring language that allows a single Web page to have different variations and to dynamically adapt itself to a given context.

2 MXML Syntax and Semantics

In an MXML document, any number of dimensions may be applied to elements and/or attributes to form *multidimensional elements* and *multidimensional attributes* whose content and value depend on their *context*. For a single multidimensional entity, contexts qualify the alternative facets of the entity. Note that, in MXML, conventional XML elements/attributes may coexist with multidimensional ones.

A multidimensional element has the form:

```
<@element_name>
  [context_specifier_1]
    <element_name attribute_specification_1>
      element_contents_1
    </element_name>
  [/]
  . . .
  [context_specifier_N]
    <element_name attribute_specification_N>
      element_contents_N
    </element_name>
  [/]
</@element_name>
```

where `element_contentsi`, with $1 \leq i \leq N$ is the content of the (multidimensional) element for the context specified by `[context_specifieri]`. The same

element name is used for all *context elements*, while the *proxy element* groups context elements and is denoted by preceding the element name with the special symbol “@”. A *context specifier* is a list of dimension assignments.

In case an attribute depends on dimensions the *attribute_specification* has the form:

```
attribute_name = [context_specifier_1] attribute_value_1 [/] . . .
                 [context_specifier_n] attribute_value_n [/]
```

In the following example, the imaginary menu of a restaurant is represented as an MXML document.

```
<menu>
  <salad vegetarian = [season = summer] "yes" [/][default] "no" [/] >
    <name> Chef's salad </name>
    <@comment>
      [language = English, detail = low]
        <comment> A traditional salad. </comment>
      [/]
      [language = English, detail = high]
        <comment>
          A salad, with a long history which
          is connected with the tradition of the town.
        </comment>
      [/]
      [language = French, detail in {low, high}]
        <comment> Une salade regionale traditionnelle. </comment>
      [/]
    </@comment>
    <@price>
      [season = summer] <price> 3 USD </price> [/]
      [default] <price> 4 USD </price> [/]
    </@price>
    <ingredient> tomato </ingredient>
    <@ingredient>
      [season != summer] <ingredient> bacon </ingredient> [/]
    </@ingredient>
    ... other salad ingredients ...
  </salad>
  ... other salads ...
</menu>
```

In this example, depending on the season of the year, the salad may contain different ingredients. Customers of the restaurant are mainly english-speaking and french-speaking. The meaning of *[default]* for a context element or attribute is that the context element or attribute holds for all the worlds for which no other context element/attribute of the same multidimensional entity holds.

For representing MXML documents we have extended *Object Exchange Model (OEM)* [1] to accommodate context-dependent information. Context specifiers

of a multidimensional entity must be mutually exclusive. That is, for a single world no more than one context element/attribute may hold in a multidimensional element/attribute. An important issue concerns the inheritance of contexts throughout the paths of the document graph. In general, contexts are inherited and further specialized as we move down a path starting from the document graph root. Note that for any world - derived by assigning a specific value to each dimension - the MXML document is reduced to the conventional XML document which holds for that world.

3 Discussion and Future Work

In this paper we briefly describe a language for representing and exchanging context-dependent information over the Web. With its simple syntax and semantics, MXML addresses a common problem in WWW information representation, namely that of describing in a robust and concise way data which appear in different variations.

Our current research directions include: (a) The formalization of semantics, and the definition of a grammar for a Multidimensional DTD. The design and implementation of a query language, tailored to handling context-specific information. (b) The possible applications of MXML seem to have a lot of potential. Using MXML to encode GIS geographical information is in our immediate plans. The representation of time-dependent information using MXML is promising, since time can be seen as an MXML dimension. A very interesting direction is the use of MXML for integration of heterogeneous information.

References

1. S. Abiteboul, P. Buneman, and D. Suciu. *Data on the Web: From Relations to Semistructured Data and XML*. Morgan Kaufmann Publishers, 2000.
2. E. A. Ashcroft, A. A. Faustini, R. Jagannathan, and W. W. Wadge. *Multidimensional programming*. Oxford University Press, 1995.
3. T. Bray, J. Paoli, and C. M. Sperberg-McQueen. Extensible markup language (XML) 1.0. <http://www.w3.org/TR/REC-xml>, 1998.
4. Brian F. Chellas. *Modal Logic: An Introduction*. Cambridge University Press, 1980.
5. M. A. Orgun and W. Du. Multi-dimensional logic programming: Theoretical foundations. *Theoretical Computer Science*, 158(2):319–345, 1997.
6. Y. Stavarakas, M. Gergatsoulis, and P. Rondogiannis. Multidimensional XML. In *Distributed Communities on the Web, Third International Workshop (DCW'2000)*, Lecture Notes in Computer Science (LNCS) 1830, pages 100–109. Springer-Verlag, 2000.
7. P. Swoboda and W. W. Wadge. Vmake and Ise General Tools for the Intensionalization of Software Systems. In M. Gergatsoulis and P. Rondogiannis, editors, *Intensional Programming II*, pages 310–320. World Scientific, 2000.
8. W.W. Wadge, G. D. Brown, m.c. schraefel, and T. Yildirim. Intensional HTML. In *Proceedings of the Fourth International Workshop on Principles of Digital Document Processing (PODDP '98)*, Lecture Notes in Computer Science (LNCS) 1481, pages 128–139. Springer-Verlag, March 1998.

AQUA

(Advanced Query User Interface Architecture)

László Kovács, András Micsik, Balázs Pataki, István Zsámboki

MTA SZTAKI

Computer and Automation Research Institute
of the Hungarian Academy of Sciences
Department of Distributed Systems

H-1111 Budapest XI. Lágymányosi u. 11. Hungary
{laszlo.kovacs,micsik,pataki,zsamboki}@sztaki.hu

Abstract. AQUA is an experimental query interface which supports iterative query refinement. Currently, it can be used as an alternative query interface for NCSTRL (Networked Computer Science Technical Reports Library) and ETRDL (ERCIM Technical Reference Digital Library). As a demonstration for the extensibility of the AQUA user interface paradigm, rating facility has been added to the system.

1 Introduction to AQUA

AQUA is a prototype implementation of a general user interface paradigm, which is capable of modelling iterative query refinement and embedding several common and advanced query techniques. The objective of this interface is not only to provide an integrated view for different searching and browsing techniques, but also to help the user to explore the contents of a digital library.

AQUA uses a single list of panels to visualize the user's query refinement steps. The first panel applies the first query operation on the whole document collection, then each panel represents an operation on the results of the previous panel. In this way users can build complex queries step by step.

AQUA currently works with the data collections of NCSTRL (Networked Computer Science Technical Reports Library) [9] and ETRDL (ERCIM Technical Reference Digital Library) [2, 4]. The screenshot on Fig. 1 shows an example query: the lower part of the AQUA screen contains a scrollable list of panels. In the example the first panel contains a boolean search for the term "digital library", with 183 results from two different digital library servers. The second panel refines the first query by searching for terms "Dienst" or "NCSTRL", and the remaining 23 documents are browsed by publication year in the third panel. The fourth panel shows the authors of documents for the years selected in the third panel. The upper part of the AQUA window serves as a console where the user can follow what is happening within the system and how their inputs are being processed. Between the console and the panel list there are tools to manipulate the whole panel list, for example queries can be saved and later reapplied to the collection.

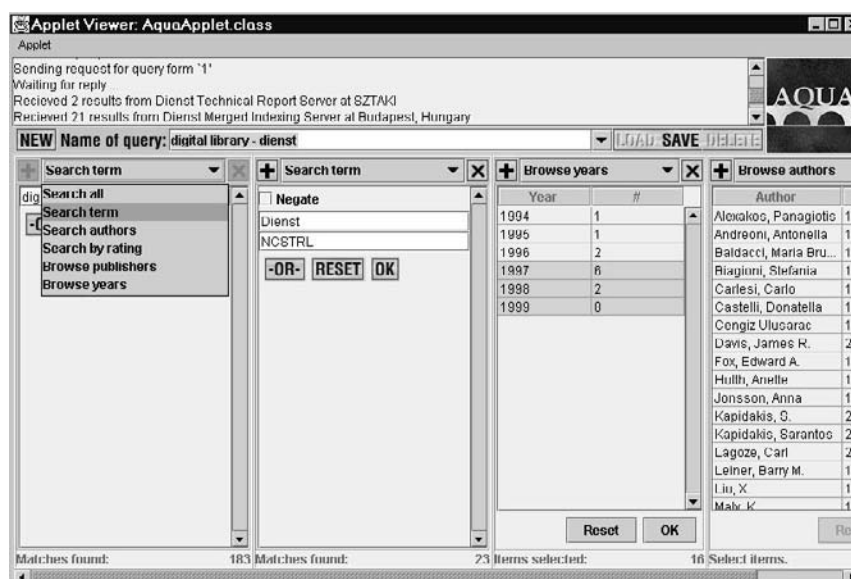


Fig. 1. A query session using AQUA

Panels in AQUA contain some standard elements: a drop-down select box at the top of the panel to change the type of query. The "delete" button in the upper right corner can be used to remove the panel from the list. The "add" button in the upper left corner is for inserting a new panel into the list before the current panel.

Both the AQUA client and the server has been implemented in Java. The client can be used as an applet, it displays the actual state of the query (e.g. Fig. 1) computed by the AQUA server. The AQUA server handles multiple client sessions, and it can access several document collections in parallel. Wrappers are used to connect different digital library systems with the server.

AQUA is available for the public since 1998 [3]. The first version was followed by a Java Swing based version in 1999. AQUA-2 has an improved user interface and provides additional features for collaborative filtering (detailed in the next section). A more detailed description of AQUA together with relations to other projects and results can be found in [5].

2 Rating Documents with AQUA

AQUA is further developed towards a collaborative tool for digital library community use. AQUA-2 helps opinion sharing and information filtering within digital document collections. This is achieved by adding rating as a new facility to AQUA-2. Users may rate documents in the collection, and this rating information can be used as other usual document metadata during queries in AQUA.

This work has been done in collaboration with the SELECT (EC Telematics Applications Programme) project [1]. The aim of the SELECT project is to provide methods for Internet users to filter the large amount of available information using collaborative techniques.

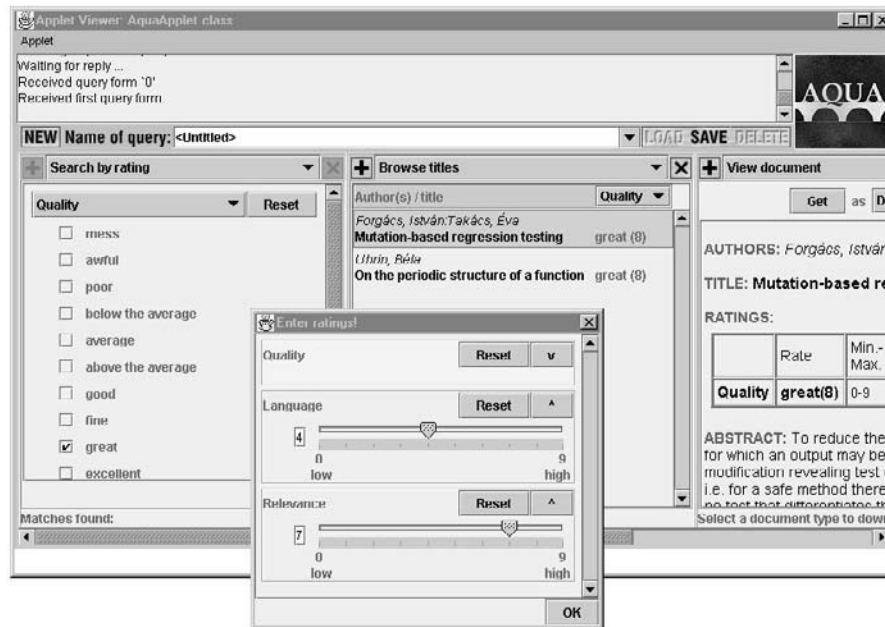


Fig. 2. Rating in AQUA

The concepts for user rating is similar to the concepts of SELECT. Users may rate each document in several categories. Categories can be defined dynamically with a given numeric scale or a set of labels. Icons and label translations for several languages can also be added to categories. User rates may have a given context. Within a context, users or user groups may have a competence value, which represents the users' expertise. Aggregate rating values for documents are calculated from rates in a single category, and presented to users. The calculation of these aggregate values may rely on competence levels, providing an architecture for many variations of rating (e.g. peer reviews, public opinions).

The user interface has been augmented with a rating input dialog, various rating displays and a new panel (Fig. 2). Aggregate rating values are shown in the *List titles* and *Document* panels. *Document* panels also show users' own rates for documents, and a new button activates the dialog for adding or changing rating values. A completely new panel type (*Search by rating*) was introduced to filter documents according to their rating values. This means that a result set can be filtered by rating values at any point in the query chain.

Currently, rating information is stored inside the AQUA system using our own implementation of rating services. In the near future it is planned to use the SELECT rating server for this purpose. This means that rating information will be globally available, not only within the AQUA system.

3 Summary

Integration of information filtering into AQUA has been presented in this paper, which is the first step towards collaboration within AQUA. In the meantime various usability evaluations are carried out to identify how well users are able to understand and apply the query chain paradigm, and what are the possible improvements of the query interface. Usability improvements and experiments with new methods for searching, filtering and collaboration characterize the evolution of AQUA.

References

1. R. Alton-Scheidl, J. Ekhal, O. van Geloven, L. Kovács, A. Micsik, C. Lueg, R. Messnarz, D. Nichols, J. Palme, T. Tholerus, D. Mason, R. Procter, E. Stupazzini, M. Vassali, R. Wheeler: "SELECT: Social and Collaborative Filtering of Web Documents and News", 5th ERCIM Workshop on "User Interfaces for All": User-Tailored Information Environments, Dagstuhl, Germany, Nov 28 - Dec 1 1999
2. A. Andreoni, M. B. Baldacci, S. Biagioni, C. Carlesi, D. Castelli, P. Pagano, C. Peters: "Developing a European Technical Reference Library", Proceedings of Research and Advanced Technology for Digital Libraries, ECDL'99, Paris, France, September 1999, Springer LNCS 1696, pp. 343-362
3. AQUA project homepage: <http://www.sztaki.hu/sztaki/aszi/dsd/aqua>
4. S. Biagioni, J. Borbinha, R. Ferber, P. Hansen, S. Kapidakis, L. Kovács, F. Roos, A. M. Vercoustre: "The ERCIM Technical Reference Digital Library", Second European Conference on Digital Libraries (ECDL'98), Heraklion, Greece, September 1998, Springer LNCS 1513, pp. 905-906
5. L. Kovács, A. Micsik, B. Pataki: "Visualizing queries in a distributed digital library system", Euromicro Workshop on Multimedia and Telecommunications at 26th Euromicro Conference: "Informatics: inventing the future", Maastricht, the Netherlands, September 5-7 2000
6. L. Kovács, A. Micsik: "The Collaborative Web", ERCIM News No. 41, April 2000, Le Chesnay ISSN 0926-4981, pp. 35-36, URL: http://www.ercim.org/publication/Ercim_News/enw41/kovacs.html
7. L. Kovács, A. Micsik, B. Pataki: "AQUA: Query Visualization for the NCSTRL Digital Library", Proceedings of the 4th ACM Conference on Digital Libraries, August 11-14 1999, Berkeley, California, pp. 230-231
8. L. Kovács: "Discovery of Resources within a Distributed Library System", Communications of the ACM, Vol. 41. No. 4. April 1998
9. B. M. Leiner: "The NCSTRL Approach to Open Architecture for the Confederated Digital Library", D-Lib Magazine, December 1998, ISSN 1082-9873, URL: <http://www.dlib.org/dlib/december98/leiner/12leiner.html>

Fusion of Overlapped Result Sets

Joaquim Macedo, António Costa, and Vasco Freitas
{macedo, costa, vf}@uminho.pt

Departamento de Informática
Universidade do Minho
Braga - Portugal

Abstract. The existence of replicated documents or indexes in distributed information retrieval introduces some level of overlapping between its component databases. In this paper, a new data fusion method taking in account the overlapping information is compared with a conventional fusion method. The overall system effectiveness evaluation enables preliminary conclusions about the importance of such parameter for data fusion in distributed information retrieval.

Keywords: Distributed Information Retrieval, Replication, Data Fusion

1 Introduction

The replication in distributed information retrieval systems is due to the existence of replicated documents in the indexed information space or even by the overlap of the indexed information space. Replication is a source of additional complexity: the text database must avoid the re-indexing of the same document; the database selection algorithm must adjust the ranking of relevant sources when a given level of relative overlapping exists; even when the query is broadcasted to all sources, the number of requested documents needs some tuning; data fusion algorithms must be able to remove replicated document items when joining different result sets.

Nevertheless, the overlapped items between different result sets may play the role of reference points, enabling an overall ranking function based on individual databases ranking. This is the main motivation for the work reported in this paper.

2 Proposed Data Fusion Method

Data fusion is not a simple problem since different databases (DBs) may use different ranking functions and/or different retrieval models. The fusion problem can be defined in the following informal way:

Definition 1. *Given a set of objects X (composed by the subsets $X_1, X_2, \dots, X_N$ with or without overlapping) and a set of ranking functions f_i (defined for the subset X_i), find an overall ranking function f defined for all elements in X .*

The data fusion problem is already known from centralized information retrieval, where it is used for enhancing retrieval effectiveness using different retrieval strategies for the same database. In this context, several fusion logics are used for computing the rank or score respecting returned document [1]. In distributed information retrieval, several merging strategies were proposed assuming disjoint databases. The main strategy used is interleaving [2].

In this work, the overlapping level between the component databases may vary from 0% to 100%. The proposed algorithm uses the conventional data fusion logics for overlapped items. From this information, the final score for unique items is estimated using a simple interpolation method. The basic idea is that scores computed by different ranking functions to the same set of documents, may be used as reference points in order to estimate the overall rank of unique items. The use of this reference points introduces some degree of comparability between non directly comparable ranking functions.

Suppose for example that there are two DBs, whose returned result sets have two overlapped documents D_1 and D_2 . f_1 and f_2 are the ranking functions for X_1 and X_2 and f is the overall ranking function for $X = X_1 \cup X_2$. Suppose that the ranking function f_i returns values between Min_i and Max_i . The first step is the combination of search engine limit values (Min_i, Max_i) to compute the limits of the overall ranking function ($NewMin, NewMax$). For instance, if the COMBSUM fusion logic is being used, the limits for the overall ranking function f will be between the lowest Min_i and $Max_1 + Max_2$.

Afterwards $f(D_i)$ is computed using one of the fusion logics for the overlapped items and using the following linear approximation for the unique ones:

$$f(u) = \begin{cases} \frac{f(D_1) - NewMin}{f_1(D_1) - Min_1} * (f_1(u) - f_1(Min_1)) + NewMin, & \text{if } Min_1 \leq f_1(u) < f_1(D_1) \\ \frac{f(D_2) - f(D_1)}{f_1(D_2) - f_1(D_1)} * (f_1(u) - f_1(D_1)) + f(D_1), & \text{if } f_1(D_1) \leq f_1(u) \leq f_1(D_2) \\ \frac{NewMax - f(D_2)}{Max_1 - f_1(D_2)} * (f_1(u) - f_1(D_2)) + f(D_2), & \text{if } f_1(D_2) < f_1(u) \leq Max_1 \end{cases}$$

The same approach is used for non overlapped items from X_2 , and generalized for m overlapped documents. Thus, this algorithm first merges the most overlapped result sets and then, if required, it performs the interleaving. The result sets are sorted by decreasing level of overlapping for the merging process.

3 Implementation and Experimental Results

The Webtrieve system [3], based on the Vector Space Model, is composed by four main components: a batch client, a broker, a search engine and a document allocator. The broker is a specialized search engine whose documents are the summaries exported by other search engines (or brokers) combined into an hierarchical information mesh. The retrieval strategy of the broker is used to select the relevant search engines for a given query. The broker has also fusion

and overlapping related mechanisms. The STARTS protocol [4] is used for communication between the different components. The actual prototype is being developed using Java and it is intended for use in Laboratory experiments with TREC collections.

The experimental results were achieved by indexing 1991 WSJ documents (TREC disk 2) with Webtrieve using stemming and stopwords remotion. To simulate the distributed environment, the documents were allocated into 32 different databases using a random allocator with replication.

In the sequel, the allocator operation is summarized. Each target database has an integer identifier between 0 and $N - 1$ (N is the number of DBs). For each document, a random value i (with $0 \leq i < N$) is generated and the document is located into the correspondent database (DB_i). For replication purposes, the algorithm needs an additional overlapping level parameter (o). The relation between o and the number of replicas (k) is given by the expression $p + \sum_{k=1}^{N-1} \frac{q}{k} = 1$, where $p = 1 - o$ is the probability of a given document d not having a replica in the overall system. The probability of d having k replicas is $\frac{q}{k}$. Thus, also for each document, a random X is generated. If $X \leq p$ the document will have 0 replicas, if $p < X \leq p + q$ one replica, $p + q < X \leq p + 1.5 * q$ two replicas, etc... After allocation, the real overlapping level between database X_i and a subset of the remaining ones may now be computed using the definition formulas or estimation methods. An overlapping level of 0.2 was used. Other experiment statistics are shown in table 1.

Collection Size (Mb)	Number of Documents	Number of Terms	Number of Postings	Number of Databases	Overlap Parameter	Average Overlap For Each DB
142,2	42 652	128 595	6 872 095	32	0.2	60%

Table 1. Statistics of the experimental scenario

A centralized database with the same data was used as the baseline on recall precision graphs. Another fusion algorithm was used for comparison [1]. Both algorithms used the COMBMAX logic. Achieved results can be seen in figure 1. Average precision values obtained are: *centralized* - 0.2595, *Fusion (WEBT)* - 0.2243 and *Fusion (old)* - 0.1901. The method using overlapping information is more effective than a conventional fusion method. For high recall values the precision is above the centralized database due to a larger number of total documents retrieved in distributed scenario.

4 Discussion

The problem of database overlapping has not been addressed in a consistent way in the literature. In the most relevant works, only brief mentions to this issue can be found [5, 2].

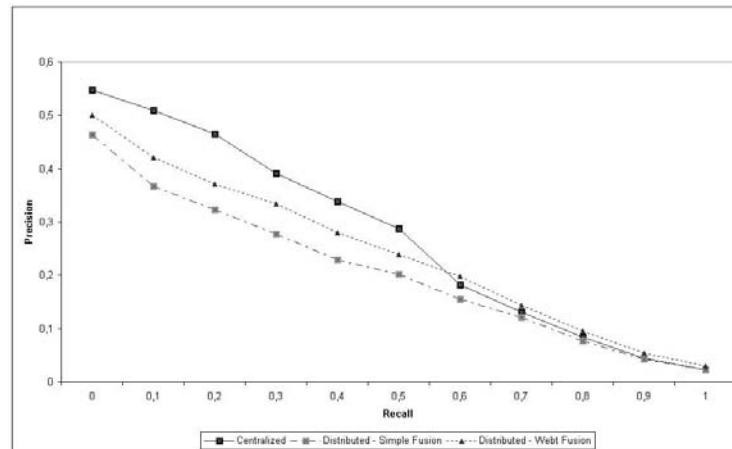


Fig. 1. Effectiveness results for the proposed fusion algorithm (WEBT)

In this work, database overlapping is considered as an important parameter to enhance distributed information retrieval system effectiveness when queried about target database contents. The fusion algorithm described here is different from those presented in the literature [5, 2]. The results presented are limited but promising. The main conclusion is that overlapping level must be used as a parameter for most distributed information retrieval algorithms.

Several issues must be addressed in future work. The first one is to get more exhaustive results for the proposed fusion method by varying the fusion logic, the overlapping level, the number of databases or other minor parameters. The fusion efficiency is also an important goal in order to merge large result sets from many databases. Another important issue is the use of overlapping information in database selection problem. The estimation of overlapping level between the databases is considered as a pre-requisite.

References

1. J. Ho Lee. Analyses of Multiple Evidence Combination. In *Proc. of ACM SIGIR Conference on Research in Information Retrieval, Philadelphia, USA*, 1997.
2. Z. Lu, J. Callan, and W. B. Croft. Applying Inference Networks to Multiple Collections Searching. Technical report, CS, Univ. of Massachusetts, 1996.
3. J. Macedo, A. Costa, and V. Freitas. Webtrieve: A distributed retrieval architecture for Text Databases. Technical report, Universidade do Minho, 1998. URL em <ftp://marco.uminho.pt/Docs/local/RT/webtrieve.ps> (in portuguese).
4. L. Gravano, K. Chang, H. Garcia-Molina, C. Lagoze, and A. Paepcke. *STARTS, Stanford Protocol Proposal for Internet Retrieval and Search*. Technical report, DL Project, Stanford University, Jan 1997.
5. N.J.Belkin and P.Kantor. Combining the Evidence of Multiple Query Representations for Information Retrieval. *IP & M*, 31(3):431-448, 1995.

ActiveXML: Compound Documents for Integration of Heterogeneous Data Sources

João P. Campos and Mário J. Silva

Faculdade de Ciências da Universidade de Lisboa, Portugal
Departamento de Informática*
{jcampos, mjs}@di.fc.ul.pt

Abstract. We address the problem of automatic composition of XML documents with data from multiple sources and their presentation to large groups of users with different information requirements.

1 Introduction

Automatic document composition by retrieving information from multiple sources faces the problem of finding a tool to seamlessly query many data types and formats. The multitude of formats of the documents to be generated presents another problem. Continuous advances in technologies for information presentation demand new formats and methods.

The database approach for information management advocates the use of declarative languages for automatic construction of documents, when retrieving data from multiple digital libraries [2]. Query languages such as Lorel [9], Quilt [3], XML Query Language [5] or XML-QL [1] allow querying an XML database for elements in XML documents, but most digital libraries today are not available as XML databases.

To cope with the heterogeneity of the existing data sources it is desirable to integrate declarative and imperative languages in one scripted document. Web servers enable the generation of documents parameterized by user input. Common Gateway Interface, Microsoft Active Server Pages (ASP) and Java Server Pages are examples of techniques that allow a web site to create a page when a user hits the page address.

Although web servers can execute multiple programs in many languages to achieve dynamic behavior, they all have the limitation that their output is of text type (usually to be interpreted as HTML by their clients). To have increased flexibility, we propose the exchange of structured data among the multiple data source handlers that could be invoked to compose dynamic web pages. With the adoption of XML we have the opportunity of developing a common structured data exchange format for heterogeneous language environments.

* This work was partially funded by research program PRAXIS XXI, medida 3.1.b)

2 ActiveXML

ActiveXML is our scripting engine for XML data integration. The operational model of activeXML is described in figure 1. It receives an XML scripted document, calls the necessary interpreters to resolve the scripted parts in the document and outputs another XML document.

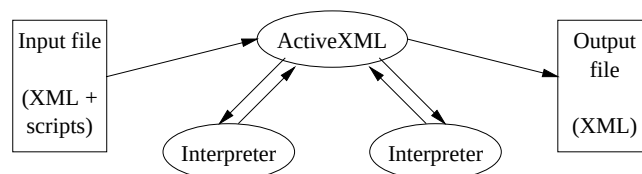


Fig. 1. ActiveXML in operation. The scripting engine receives a document with embedded scripts as input. It runs through the document and copies every element to the output document. If the element to be copied is a `<script>` element, it calls the interpreter indicated as an attribute along with the XML element as input. The interpreter then follows an established policy for handling the contents of the script elements and returns a well formed XML element that the scripting engine then appends to the output document

ActiveXML only outputs XML data, but XML is capable of representing most data structures. In addition, as it can be integrated with most other XML-based software, such as XSLT (eXtensible Stylesheet Language Transformations) [13], it can easily be extended to generate any kind of textual data. Our approach for supporting multiple interpreters for data integration from many sources enables the use of all languages proposed for handling web data sources. ActiveXML differs from mediators [4] in its approach for data integration. Mediators receive queries in a unified query language, translate them into the query method of the data source and return result data; ActiveXML receives queries in any language, invokes the interpreter for that language to handle the query and collects the output as XML data. This multi-language approach enables the use of the most expressive language available for each task.

3 Online Web News Personalization with ActiveXML

We developed ActiveXML in the context of a daily online newspaper [8, 6] that needs to publish multiple front pages generated, for a set of user profiles. The front pages contain news from the newspaper archive and other sources, such as stock quotes for the “Business” user profile or the weather forecast and movie listings for the Lisbon local news. The front pages have to be automatically generated and kept up to date. To specify the front page for a given profile, the editor selects the three most important news, the main stock quotes and two

targeted advertisements. This page is then published in HTML for the Web, as WML for WAP devices [14], and as data to be retrieved by dedicated browsers in Palm Pilots [7].

In this environment, ActiveXML is at the core of the data retrieval/generation processes. To gather information from the sources, a single script produces a document with all the data to be published (see figure 2(a)). In the next stage, we process the documents with the news of each profile to be transformed for publication in the supported media. As the documents are in XML, we use XSLT to translate them into the desired formats. As non XSLT formats might be needed (like Palm Pilot's database streams) we replay the documents through ActiveXML, but this time the interpreters have the side effect of generating new documents in specialized formats (see figure 2(b)).

This approach makes the various steps of the document generation process independent. The contents update policy specification is separated from the document to enable easy re-generation of customized front pages. The document contents generation is separated from its style specification to enable changing the style while keeping the documents structure. This also encourages reusability. In our environment, the WebGenerator, which outputs static HTML pages based on an XSL stylesheet, can also produce WML contents, given the appropriate stylesheet.

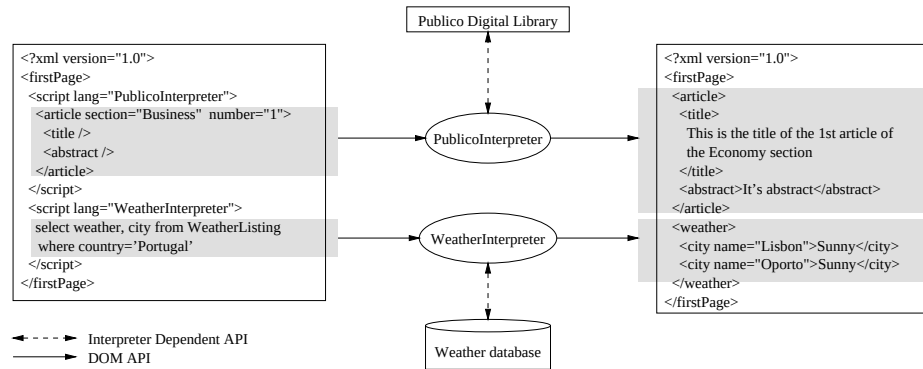
Although ActiveXML aims at integrating and publishing information, the virtual versus materialized approach for data integration is left open to the developer of the interpreters and generators. For example, in the case of the publication of daily news one might want to materialize the data, to obtain the best performance with the web server. Our WebGenerator interpreter when processing stock quotes outputs code for a script enabled web server to get the data on the fly.

4 Conclusions and Future Work

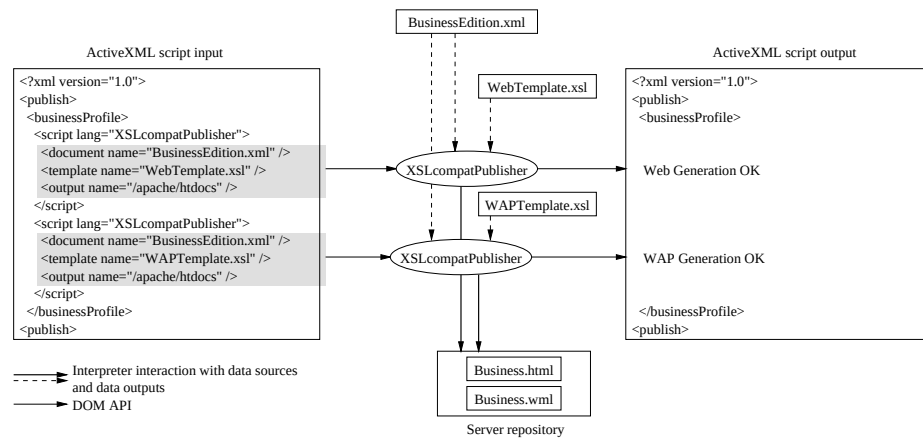
We developed a framework for integrating XML data and programs and process them in a structured and uniform way. This is achieved through a model for the publication in multiple formats of composite documents made from information items coming from multiple sources.

ActiveXML is very general and flexible but has many limitations. Among others it is hard to manage link relationships between generated documents. However some of the interpreters, such as JavaScript and VBScript can manipulate XML documents through a common API (DOM [11]). This could be used in the future for sharing data among interpreters and ActiveXML.

ActiveXML will be one of the components of XMLBase [10], our new component based system for XML data management, under development.



(a) Composing an XML document with activeXML. In our prototype, the script invokes two interpreters that query two data sources: *PublicoInterpreter* retrieves articles from *Publico's Digital Library*; *WeatherInterpreter* queries an SQL weather database



(b) Handling multiple formats: the script specifies the output formats intended for the document. The activeXML engine successively calls interpreters to retrieve the data document and a stylesheet. It generates a log of its actions as an XML file and, as a side effect, data for user presentation is saved in a repository, ready to be served to clients

Fig. 2. ActiveXML in action, generating the front pages for a personalized newspaper in multiple formats

References

1. Alin Deutsch, Mary Fernandez, Daniela Florescu *et. al.*. A Query Language for XML. <http://www.research.att.com/~mff/files/final.html>, May 2000.
2. Daniela Florescu, Alon Levy and Alberto Mendelzon. Database Techniques for the World Wide Web: A Survey. In SIGMOD Record 27(3), 1998.
3. Don Chamberlin, Jonathan Robie and Daniela Florescu. Quilt: An XML Query Language for Heterogeneous Data Sources. In WebDB 2000, Proceedings of the Third International Workshop on the Web and Databases, May 2000 (to be published as a Springer LNCS volume).
4. Gio Wiederhold. Mediators in the Architecture of Future Information Systems. IEEE Computer, March 1992.
5. J. Robie, J. Lapp, D. Schach. XML Query Language. <http://www.w3.org/TandS/QL/QL98/pp/xql.html>, May 2000.
6. Nuno Maria, Pedro Gaspar, Nuno Grilo, António Ferreira, Mário J. Silva. ARIADNE - Digital Library Architecture. In Proceedings of the 2nd European Conference on Digital Libraries (ECDL'1998), 1998.
7. Palm Inc. Handheld Computing Devices. <http://www.palm.com/>. May 2000.
8. Público Online. <http://www.publico.pt>. May 2000.
9. Serge Abiteboul, Dallan Quass, Jason McHugh, Jennifer Widom and Janet L. Wiener. The Lorel Query Language for Semistructured Data. International Journal on Digital Libraries, 1(1):68-88, April 97.
10. XMLBase - semi-structured data management system. <http://xldb.fc.ul.pt/xmlbase/index.html>, May 2000.
11. W3C. Document Object Model. <http://www.w3.org/DOM/>. May 2000.
12. W3C. XML Linking Language (XLink), W3C Working Draft 21-February-2000. <http://www.w3.org/TR/2000/WD-xlink-20000221/>
13. W3C. XSL Transformations (XSLT), Version 1.0. W3C Recommendation 16 November 1999. <http://www.w3.org/TR/1999/REC-xslt-19991116>
14. Wap Forum. <http://www.wapforum.org/>. May 2000.

newsWORKS©, the Complete Solution for Digital Press Clippings and Press Reviews: Capture of Information in an Intelligent Way

Begoña Aguilera Caballero¹ and Richard Lehner²

Parc Tecnològic del Vallès. Centre d'Empreses de Noves Tecnologies, of. 2
08290 Cerdanyola (Spain)

¹ baguilera@inicia.es

² Richard.lehner@t-online.de

Abstract. A new software solution is presented, specially designed for the electronic handling of press clippings, in order to build press archives and to produce press reviews in a digital way. Based on the ultimate standard technology available on the market, newsWORKS© is able to automate the layout analysis of the newspaper and the recognition of the articles. It offers too, the best OCR tools, besides with manual tools to add the intellectual work that at the end has to be made by the specialists (intellectual indexing).

1 Introduction

Information published in newspapers and magazines is absolutely relevant for most of the users of Information, documentation centres and libraries around the world, as well as for marketing and press departments of companies of different sizes. Their users are not only interested in the content of those articles but also in the typography, the layout, the pictures that come with it, the impression of the whole page. For this reason, these centres produce press clippings that will be either archived for future access, and in some cases, producing press reviews to distribute among its users.

Tools used by these centres are even today, mainly the analogical ones: scissors and glues. Sources are photocopied, cut, pasted and copied again as much times as necessary; either for archiving under different indexes and/or for distributing as press reviews. Some software is available in the market yet, offering a manual solution, a hybrid between the analogical world and the automated solution we are going to present. In these cases, the workflow would be to scan without cutting the article interested, and then to “crop” manually with the keyboard or the mouse the surface of the article. Then, manually too, the indexing is added by the user.

The innovative system we present, newsWORKS© allows professionals working with clippings, to produce –without modifying the internal workflow- press clippings and press reviews in a fast, accurate and intelligent way.

2 Main Concepts and Features

When producing clippings in a digital way, some main points arise, that are easily solved by the system solution we present.

Recognition of Selected Articles.

It is very important, when working with documentation, to recognize and differentiate what is a non-important task, routinely, and what is an added-value task, that is worthwhile the effort a person will do. When producing clippings, for example, a routinely task is the recognition of the zone occupied by the selected articles, and the goal should be to automate as much as possible this task. Better to use the time of the staff in another added-value tasks.

The module for clipping of newsWORKS®, newsCLIP®, is able to automatically recognise text and images zones that belong to an article, and assemble them accordingly in a target page [see fig.1] The automatic layout analysis is able to select the articles on a page as well as textual objects within the article, that is title, subtitle, author, abstract, body text, picture and caption.



Fig. 1. The most interesting ability of newsCLIP®: automatic layout analysis of the source and recognition of objects.

Recognition of articles and objects inside the articles is successfully done even with polygonal zones that present an added difficulty for automated systems.

Production of the Clipping

The chosen article should be fitted into a target page, considering the original size, which is important to be able to evaluate further the importance of the article. newsCLIP© is able to paste the clipping into the target page, with just one click. Articles are automatically arranged in an A4 page that contains the thumbnail of the source page

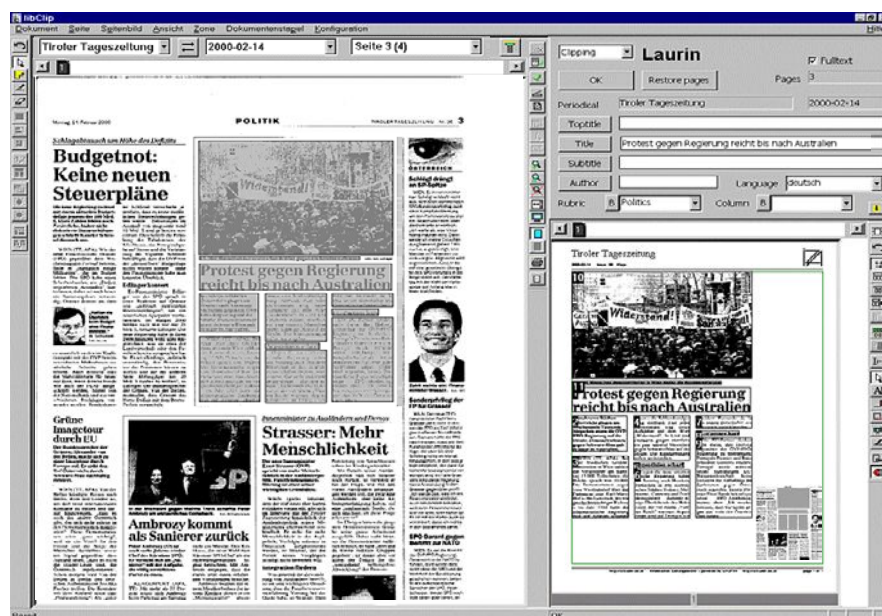


Fig. 2. Detail of the target page, with the clipping and the thumbnail of the complete newspaper page

These two first points (recognition of selected articles and production of the clipping) must be done in the easiest way possible to minimise the resources used for producing clippings. With newsWORKS the procedure has been reduced to three clicks of the mouse: one for recognising the article and its zones, and one for the clipping.

Primary and Bibliographic Indexing (Metadata)

To be able to identify the articles is mandatory to produce a primary indexing, and for further retrieving from the database where they will be exported. With newsCLIP©, at the same time that the “pasting” of the article is done in the target zone, the indexing of defined zones (title, author, etc.) is automatically done, with the use of OCR technology. With the same effort (cutting and pasting) users get a first added value: basic indexing by the fields previously defined by them.

Full-Text Recognition

Three outputs of the clipping process will be exported to the database –if the user is interested in three of them-. We have seen how newsWORKS© is able to produce the electronic facsimile of the article (first output) and the prime and bibliographic indexes (second output). The third output will be the OCR-processed full text of the article.

In the market there is no OCR engine able to get a 100% failure free text. In practical, this means that very precious time and effort is lost correcting OCR results. newsWORKS uses the concept of newsREAD/Voting OCR© to improve the performance of these engines. Three OCR software packages –at the moment Textbridge, Omnipage and FineReader- work simultaneously in the same text, ensuring that the correct character is detected. Not only the attributes of the individual characters are analysed but also the characters are examined in a linguistic context via lexical field analyses. This process assures an increase of the performance of OCR technology of a 30%.

Creation of Press Reviews

Assembling and arranging topical recorded articles in a press review can be tiring work and also takes up a great deal of time. Requirements for saving time and effort are that the system must be easy to select the articles and also a clearly arranged press review must be guaranteed. With newsPRESS© specific items are selected from the available articles using different criteria and are deposited in individual selection lists. Articles are assembled automatically on the pages of the press review and arranged so that they fit perfectly. List of contents and cover page for each press review is automatically produced. The press review can then be printed and saved as TIFF data, Adobe Acrobat document or even published in HTML format.

3 Conclusion

newsWORKS© is the system specifically designed for press clipping, that allows clippers to do their job in a very fast and accurate way, with the added value of building an electronic archive of clippings, with no additional effort and as many press reviews as necessary. newsWORKS© is the innovative electronically solution for the automated clipping production. We are talking about clippings in seconds, even with full-text, indexing and distribution over Intranet: the innovative way of contemporary clipping.

Effects of Cognitive and Problem Solving Style on Internet Search Tool

Tek Yong Lim and Enya Kong Tang

School of Computer Sciences
Universiti Sains Malaysia
11800 Penang, Malaysia
tylim@cs.usm.my, enyakong@cs.usm.my

Abstract. This paper presents a research proposal on user-oriented evaluation method to compare the usability of Internet search tools. Cognitive style and problem solving style are identified individual difference factors. Meta-search, portal and individual search engines are Internet search tool available. Usability of each search tools based on relevancy and satisfaction is another factor of this study. The ultimate aim of the research is to contribute to the knowledge concerning individual differences and information retrieval technology. In particular we hope to get a better understanding of which presentation structures and user interface attributes work best and why.

1 Introduction

Searching for information through the Internet browser is widely used by different level users. Currently, the Internet search tools are becoming more complex and even more frustrating for users. According to Leader and Klein (1996), users with different cognitive styles tend to develop and use different strategies in a search interface. Extensive research had been carried out to measure each Internet search tools performance and interface design. Most of these findings were not based on actual tests and measurements. Most of the researches rather than users make relevance judgements on the Internet search tools [1], [2], [5], [18], [19]. Research investigating how different users use a search tool and identifying the factors affecting the usage is important to help developing user-friendlier search tool [20]. Previous studies of searching behavior suggest that differences in individuals search strategies, effectiveness of searches and the satisfaction with the result of searches are significantly linked to the differences in cognitive style [6], [16]. This study aims to examine such findings in the context of Internet searching tools.

This study proposed to find how the user use the Internet search tools when searching information and how the users' psychological factors including cognitive and problem solving styles influences the searching patterns. Besides that, possible effects of type of search tasks will be investigated too. We will look into the possibility of involving users as evaluator on evaluating the efficiency and effectiveness of Internet search tools. In this paper, we will try to look for the following question:

- What is the effect of user's cognitive styles on the information search strategies on the Internet search tools?
- What is the effect of user's problem solving style on the information search strategies on the Internet search tools?
- Do types of search task influence the user's information search performance and strategies?
- How well does Internet search tools perform from end-user judgements?
- Do users perceive interface design to be important when using Internet search tools?
- What kinds of interfaces for search engines or outputs provide maximum interaction between the system and the user?

In section 2, we discuss the type of search tools available on the Internet, individual differences and usability. In section 3, we review the related studies in the past. Conclusion is presented in Section 4.

2 Definition of Terms

2.1 Internet Search Tools Differences

Internet search tools have been divided into different type based on their unique interface and services they provided over the Internet. We have identified three types: Individual search engine, Portal and Meta-search engine.

Individual Search Engine (ISE). Search engine is a program that searches through some dataset. Individual search engine run search algorithms based on user-input text expressions in their database. It presents a simple input textbox and a push button on the search page. The interface looks simple and easy to use. Users will have to formulate their query using appropriate search language like keywords and/or phrases with Boolean operator. They can also use parentheses and standard predicate calculus precedence to make a precise and unambiguous query. Then users will have to search through a set of results. The simple result page will display top 10 result retrieved from their indexed database. It does not have any other features on the result page. These engines will have a tendency to retrieve more duplicates, linkrots and plagued by low precision.

Portal (P). Portal is a doorway, an entrance or a gate, especially one that is large and imposing. Portal is a new term, generally synonymous with gateway, for a World Wide Web site that is or proposes to be a major starting site for users when they get connected to the Web or that users tend to visit as an anchor site. Typical services offered by portal include a directory of Web sites, a facility to search for other sites, news, weather information, e-mail, stock quotes, phone, map information, and a community forum. Most of them provide user with too many option and feature and the interface looks a bit cluttered. The reason is that they want to keep their users surfing inside their portal. Sometimes users do not even know where and how to search. Users may start their search using simple search textbox or using advanced

search. The result page will provide many refinement features to help the user to narrow down their search like category matches, related search and multimedia. They even can begin their search through directory services. They may end-up frustrated and give-up when they cannot find any information [21].

Meta-Search Engine (MSE). Meta-search engine allows a user to submit a query to several different search engines for searching all at once. For this paper, the focus of the meta-search engine is browser add-on search tools. Add-on is one thing added as a supplement to another, especially a component that increases the capability of the system to which it is added. Copernic 2000, Web Ferret, Beeline and Quest 99 are among the browser add-on search tools. Since it has its own interface, users need to switch between two windows (web browser and meta search engine) when searching for information. Results are merged with duplicates and linkrots removed. However, it may takes longer time to process user query. Navigation between browsers and add-on tools need to be considered also.

2.2 Individual Differences

Human approach mental tasks with different ways of perceiving and thinking, and this is determined by their cognitive styles which explains the variations in modes of perceiving, remembering, and thinking, or, if the information processing framework is utilized, as distinctive ways of apprehending, storing, transforming and utilizing information. It seems reasonable to argue that different individual will have different strategies when processing information. We will look into cognitive and problem solving style in individual differences.

Cognitive Style. Cognitive style is known as a tendency for an individual consistently to adopt a particular type of strategy. Cognitive style also refers to a manner of moving toward a goal and a characteristics way of experiencing or acting. It is the characteristic way in which the individual organizes and process information. Cognitive style can be measured in several different dimensions. The cognitive style of interest in the present study is the extent to which a learner is field dependent or independent. Field dependent is a dimension of individual difference that extends across perceptual and intellectual functioning [22]. Relatively field-dependent and field independent persons tend to favor different learning approaches. While the internal variable of interest in the present study is the cognitive style of field dependents/independents, the external variable of interest is the different interface of presentation of the Internet search tools.

Problem Solving Styles. Problem solving can be defined as a goal oriented sequence of cognitive operation. The problem solving process comprises cognition as well as emotion and behavior. Skills of problem solving include the ability to search for information, to analyze situations for the purpose of identifying the problem in order to generate alternative courses of action, to weigh alternative courses of action with respect to desired or anticipated outcomes, to select and implement an appropriate plan of action and to evaluate the outcome with reference to the initial problem. Problem solving style in this study is defined as a tendency to respond in a certain

way while addressing problems and not as the steps employed in actually solving problem. Two problem solving styles defined in this study are low and high Problem Solving Inventory style [9].

2.3 Usability

Usability had become an issue in information retrieval and World Wide Web. Most of the reviewer evaluation on the information retrieval tools will have their own checklist [3], [4], [7], [10]. Checklists tend to be a tool for reviewers rather than end-users. Reviewers may talk of systems being 'user-friendly' or having poor feedback, or slow and user may be left wondering how this has been measured or whether user would agree with it. Therefore, there is a need for modification on the criteria in order to let the user become evaluator for Internet search tools (refer to Table 1).

Table 1. Usability of Internet search tools

Criteria	Measurement
Relevance	Precision Number of relevant document Relative recall
User Satisfaction	Response time Search interface Online documentation Output format Overall Reaction

3 Related Studies

Many researchers have considered Internet searching to be one of the most challenging and rewarding areas of research for future information retrieval application. Because not all users will search the Web in the same way, individual differences may cause difficulties in using the Web to find information. Henninger and Belkin (1996) stated that information retrieval research can be divided along the lines of its system based and user based concern. The user-based view must account for the cognitive state of the searcher and the problem-solving context. There are few studies that have investigated cognitive style as a factor in use of hypermedia systems found performance differences between field independents and field dependents [12], [17], [14]. Nahl and Tenopir (1996) have demonstrated the importance of the affective domains as complements to the cognitive elements of online searching behavior. Ford, et al (1994) found significant correlations between cognitive style and online searching (LISA CD-ROM). Leader and Klein (1996) also revealed that a significant interaction between search tool and cognitive style in hypermedia database search.

Individual difference studies on Internet have been done but some of the previous studies that had been carried out are either unsystematic or unclear measurement. The measures were not always defined clearly and measures with the same names might

be defined and implemented differently [11], [15], [20]. Problem solving is another cognitive process required for using an information retrieval system. Problem solving starts with a perceived problem. Once the problem is stated and understood, individuals then apply their knowledge to the problem and attempt to try out possible solutions. Solutions obtained are evaluated with reference to initial problem definitions. How the problem solving process affects the Internet search is an area worthy of being studied.

4 Conclusion

This study aims to examine such findings in the context of Internet searching tools. We will try to investigate the effects of cognitive and problem solving style on performance in Internet search. This paper will collect and produce qualitative and quantitative analysis to provide a better understanding of the Internet searching. Statistical analysis of quantitative and content of analysis of verbal data will help to establish a body of sound knowledge about the user-oriented evaluation. Leader and Klein (1996) expected that the treatment type of search tools would interact with the cognitive styles of the users. Besides that, we will look into usability of each search tool from the user perspective. By understanding the effects of different individual cognitive and problem solving style on Internet search, we can provide a better understanding the individual strategies used for different types of search interface and improved the development of Internet search tool systems. In particular, we hope to get a better understanding of which presentation structures and user interface attribute work best and why.

References

1. Baker, A.L.: Datastar Web: A Comparison with "Classic" Datastar Command Language Searching. Online & Cdrom Review, (1998) 22 3
2. Chu, H.T., Rosenthal, M.: Search Engines for the World Wide Web: A Comparative Study and evaluation methodology. Proceeding of the American Society for Information Science (1996)
3. Courtois, M. P., Baer, W. M., Stark, M.: Cool Tools for Searching the Web, Online, (1995) 19 6
4. Desmarais, N.: The Librarian's Cd-Rom Handbook, Meckler Publishing, London (1989)
5. Ding, W., Marchionni, G.: A Comparative Study of Web Search Service performance, Proceedings of the American Society for Information Science (1996)
6. Ford, N., Wood, F., Walsh, C.: Cognitive Styles and Searching. Online & Cdrom Review, (1994) 18 2
7. Harry, V., Oppenheim, C.: Evaluations of Electronic Databases, Part 1: Criteria for Testing Cdrom Products. Online & Cdrom Review, (1993) 17 4
8. Henninger, S., Belkin N.J.: Interface Issues and Interaction Strategies for Information Retrieval Systems. Proceeding of SGICHI (1996)
9. Heppner, P.P.: The Problem Solving Inventory. Consulting Psychologists Press (1988)
10. Herther, N.: Text Retrieval Software For Microcomputers, Online, (1986b) 10 5
11. Hsieh Yee, I: Search Tactics of Web Users in Searching for Texts, Graphics, Known Items and Subjects: A Search Simulation Study. The Reference Librarian, (1998) 60

12. Jonassen, D.H., Wang, S.: Acquiring structural knowledge from semantically structured hypertext. *Journal of Computer Based Instruction*, (1993) 20 1
13. Leader, L.F., Klein, J. D.: The Effects of Search Tool Type and Cognitive Style on Performance During Hypermedia Database Searches. *Educational Technology Research & Development* (1996) 44 2
14. Liu, M., Reed, M.W.: The relationship between the learning strategies and learning styles in a hypermedia environment. *Computer in Human Behavior*, (1994) 10
15. Moss, N., Hale, G.: Cognitive Style and Its Effect on Internet Searching: A Quantitative Investigation. *Proceeding of European Conference on Educational Research*. (1999)
16. Nahl, D., Tenopir, C.: Affective and Cognitive Searching Behavior of Novice End-Users of a Full-Text Database. *Journal of the American Society for Information Science* 1996 47 April
17. Repman, J., Rooze, G. E., Weller H.G.: Interaction of learner cognitive style with components of hypermedia-based instruction. *Journal of Hypermedia and Multimedia Studies*, (1991) 2 1
18. Su, L.T., Chen, H.L.: User Evaluation of Web Search Engines As Prototype Digital Library Retrieval Tools. *Proceeding of 3rd International Conference on Conceptions of Library and Information Science* (1999)
19. Tegenbos, J., Nieuwenhuysen, P.: My Kingdom for an Agent? Evaluation of Autonomy, an Intelligent Search Agent for the Internet. *Online & Cdrom Review*, (1997) 21 3
20. Wang, P., Tenopir, C.: An Exploratory Study of Users' Interaction with World Wide Web Resources, Information Skills, Cognitive Styles, Affective States and Searching Behaviors. *Proceeding of 19th Annual National Online Meeting*. (1999)
21. White, M. D., Iivonen, M.: Factors Influencing Web Strategies. *Proceeding of ASIS Annual Conference* (1999)
22. Witkin, H.A., Moore, C.A. , Goodenough, P.R., Cox, P.W.: Field dependent and field independent cognitive styles and their educational implications. *Review of Educational Research*, (1977) 47

Follow the Fox to Renardus: An Academic Subject Gateway Service for Europe

Lesly Huxley¹

¹Institute for Learning and Research Technology
University of Bristol, 8-10 Berkeley Square, Bristol BS8 1HH, UK
lesly.huxley@bristol.ac.uk

Abstract. Renardus is a collaborative project of the EU's Information Society Technologies programme with partners from national libraries, university research and technology centres and subject gateways Europe-wide. Its aim is to build a single search and browse interface to existing quality-controlled European subject gateways. The project will investigate related technical, information and organisational issues, build a pilot system and develop a fully-operational broker service. This paper provides an overview of the project, work in progress and anticipated results and outlines the opportunities and benefits for future collaboration in developing the service.

1 Overview

As the Internet continues to expand it is clear that no single, publicly-funded subject or national gateway/digital library initiative can hope to identify, catalogue and organise all the Internet resources available to support Europe's academic and research communities. Renardus is a collaborative project of the EU's Information Society Technologies programme with partners from national libraries, university research and technology centres and subject gateways. Their aim is to build a single Web service to search and browse across existing European scientific and cultural resource collections. Between January 2000-June 2002 the project will: investigate related technical, information and organisational issues; build a pilot system in 2001 (to be verified through addition of at least one more gateway); establish a testbed environment for experimentation with data sharing and multilinguality issues and, in June 2002, develop a fully-operational 'broker' service.

Project partners [1] from across Europe are together addressing the considerable implications for technical and information standards, business and sustainability issues. The potential benefits of collaboration for participating service providers include: scale economies in metadata creation, abstracting and indexing; a more sustainable level of quality in mediated resource discovery; improved sustainability and a stronger position against international competition. Full project reports, summaries and links to related information are available from the Renardus Web site [2]. Completed work and work in progress are summarised below.

2 Surveys of User Requirements and Use Case Scenarios

User requirements have been collected at various levels to inform the service's functional specification and design. Respondents to a survey of service providers favoured a distributed architectural model: the concept of a central repository to which all metadata is routinely copied was rejected in favour of a centralised subject index to forward queries to relevant gateways. Support was also required for Dublin Core semantics and RDF/XML syntax for metadata records; Z39.50 and WHOIS++ communication protocols and mappings between metadata formats to support a consistent presentation of search results. Mappings between classification schemes (even if only at the highest level) are also needed for implementation of the cross-browsing functionality in the pilot system. Full details and technical references are provided in the report *User Requirements for the Broker System*. [3].

Partners also gathered data from a range of end user surveys previously undertaken by participating gateways. Generally, these showed that users are more at ease with navigating the services gateways offer than is the case with the 'rest' of the Internet. Users were found to appreciate the evaluation, categorization and currency of resources offered and to prefer simple searches with preferred categories including keyword, author, title and description.

Use case scenarios and activity diagrams using the Unified Modeling Language [4] have been developed to describe how various players will use the service, without specifying any technical solutions for how this will be achieved. The use cases build on the requirements outlined in the user and service provider surveys. The individual use cases are intentionally narrow in focus, covering functions such as: simple searching, cross-browsing by subject or displaying results. Other use cases cover administrative functions for maintaining metadata indexes, assuring data quality, or adding data.

3 Evaluation of Broker Models in Related Projects

One of the biggest challenges for those currently developing digital libraries is how to provide integrated access to the wide range of distributed and heterogeneous information resources and services available. The success of this integration is seen as beneficial both to libraries and their end users. Dempsey, Russell and Murray (1999, p.35) [5] recognise that users may often have to negotiate several quite different information systems and interfaces to complete a full search. They suggest development of an additional service layer ('middleware') which, in our context, is the Renardus broker service. A comprehensive Evaluation of Existing Broker Models in Related Projects [6] undertaken by Renardus partners provides a map against the generic model known as the MODELS Information Architecture (MIA) [7]. As a similar review was being undertaken at the same time for the IMesh Toolkit project [8], some of the latter's reviews have, with kind permission, been adapted and included for Renardus' review.

4 Service Scope, Data Model, and Technical Standards

The scope of the pilot system (and objectives for the fully-operational service) have been agreed, including subject and geographical coverage, definitions and criteria for participating gateways. The Scope Document [9] includes terms and definitions of concepts based on those defined in the DESIRE Information Gateways Handbook [10] and the overview of gateways provided by Koch, 2000 [11]. A number of new concepts are also introduced including “open subject gateways” and “resource discovery broker services”. Renardus partners have agreed an initial set of thirteen starting points for the pilot system: the intention is to refine these as the project progresses.

A Review of Existing Data Models [12] has led to the development of a minimum common set of metadata elements for Renardus as a first step towards developing the service’s data model. In response to questionnaires, gateways provided details such as collection description, target group, resource categories, quality criteria, controlled vocabularies and descriptions of their respective metadata sets. Based on these, a metadata mapping has been developed to produce a minimum common set of metadata elements. The data model will continue to be developed and refined, but currently includes: DC.Title; DC.Creator, DC.Description, DC.Identifier, DC.Subject, DC.Publisher, DC.Language, DC.Type.

A Review of Technical Standards and Solutions [13] in the areas of information retrieval and searching considers potential for use in Renardus. Standards chosen for the review include: Dublin Core, CIP, HTTP, IAFA, LDAP, RDF, XML, Z39.50, Z39.50 to Web Gateway, and WHOIS++. A short summary is provided for each standard, covering functionality, strengths and weaknesses and relevance to Renardus, with links to more detailed information available elsewhere. The result is a useful overview of the technical standards likely to be encountered by anyone involved with subject gateways and Web information retrieval.

5 Further Information and Opportunities for Participation

Dissemination of project results is being undertaken alongside the research and development work through the Renardus Web site and a regular email newsletter (the Renardus News Digest [14]), conference papers and presentations. Full reports and summaries of project findings are offered on the project Web site as soon as they are available, together with descriptions of related projects and services across Europe and beyond. A workshop for potential participating services is to be held in September 2001. Identification of potential participants is underway. User guidelines on data interoperability will be developed to facilitate participation in the Renardus broker service.

Renardus partners are particularly keen to enter dialogue with other organisations that may want to participate in the verification of the pilot service or in the fully-operational broker service. A feedback form [15] is available on the Web site to facilitate this and provide a route for news and information from other initiatives.

6 Conclusions

Our work so far and contacts with related projects and services has supported our initial view that collaboration at a European level is both feasible and potentially beneficial both to service providers and end users alike. The collaborative framework emerging from the Renardus project will increase expertise in the provision of quality and sustainable subject-services within the many participating services. Strong collaborative impulses are already coming from the United States and Australia, trying to extend collaborative projects to Europe. Within the Renardus collaborative framework there is potential for positioning European participants more strongly on the international scene and putting them in an advantageous position to seek partners for international collaboration.

References

1. Partner organisations are described on the Web site page:
URL: <<http://www.renardus.org/about/partners.html>>
2. Renardus Web site home page
URL: <<http://www.renardus.org/>>
3. Internal project deliverable "User Requirements for the Broker System"
URL: <http://www.renardus.org/deliverables/#D1.2>
4. Unified Modelling Language (UML) described in Martin Fowler, with Kendall Scott, 2000, UML Distilled Second Edition: A Brief Guide to the Standard Object Modelling Language. Addison-Wesley
5. Dempsey, L., Russell, R., Murray, R., 1999, A utopian place of criticism? Brokering access to network information. *Journal of Documentation*, 55 (1), 33-70
6. Public deliverable "Evaluation of Existing Broker Models in Related Projects"
URL: <<http://www.renardus.org/deliverables/#D1.1>>
7. MODELS - MOving to Distributed Environments for Library Services
URL: <<http://www.ukoln.ac.uk/dlis/models/>>
8. IMesh Toolkit project
URL: <<http://www.imesh.org/>>
9. Public project deliverable "Scope Document"
URL: <<http://www.renardus.org/deliverables/#D6.2>>
10. DESIRE Information Gateways Handbook
URL: <<http://www.desire.org/handbook/>>
11. Koch, T (2000), Quality-controlled subject gateways: definitions, typologies, empirical overview, Subject gateways special issue of *Online Information Review* Vol. 24:1, Feb 2000, MCB Univ. Press, available at:
URL: <<http://www.lub.lu.se/~traugott/OIR-SBIG.txt>>
12. Internal project deliverable "Evaluation of Existing Data Models"
URL: <<http://www.renardus.org/deliverables/#D6.1>>
13. Internal project deliverable "Technical Standards and Solutions"
URL: <<http://nwi.dtv.dk/RENARDUS/D2.1>>
14. Renardus News Digest
URL: <<http://www.renardus.org/news/digest.html>>
15. Feedback form for all contact with Renardus project partners
URL: <<http://www.renardus.org/contacts.html>>

CORC: Helping Libraries Take a Leading Role in the Digital Age

Kay Covert

OCLC Online Computer Library Center, Inc.
covertk@oclc.org

Abstract. The OCLC Cooperative Online Resource Catalog is helping librarians thrive in the digital age. Librarians are using CORC to select, describe, maintain, and provide guided access to Web-based electronic resources. Librarians in more than 24 countries are using CORC and all types of libraries, including public, academic, corporate, school, and government libraries are contributing records to the CORC catalog. The CORC service offers a Web-based toolset for cataloging electronic resources, a robust database of high-quality resources, and a tool for building dynamic pathfinders. Developed by and for librarians, CORC blends three key elements—technology, cooperation and librarianship, to help librarians define the future of knowledge access management.

Toolset for Selecting, Describing, and Maintaining Electronic Resources

Using the CORC automated tool kit, librarians point the CORC system to a URL and CORC extracts data from the resource to create a record. CORC's underlying technology uses XML/RDF templates to drive a Web-browser-based resource description editor on top of a standard OCLC SiteSearch software database. This software allows CORC to support resource description work in both MARC and Dublin Core formats concurrently, and will allow CORC to accommodate other metadata formats such as TEI and EAD. The software architecture also allows for online linked authority files as well as links for Dewey Decimal classification of resources, OCLC name and corporate name authorities, and Library of Congress subjects. The resource description system provides further automation through librarian-directed web site harvesting and author-supplied metadata capture as well as automatic subject assignment¹. Completed bibliographic records can be exported in either MARC21 or Dublin Core XML/RDF or Dublin Core HTML formats to the library's local system.

With CORC, member libraries have the added advantage of shared URL maintenance. CORC provides URL checking and notifies librarians who have

¹ Taylor Surface, CORC: Library-directed Selection, Description, and Access to Web and Electronic Resources, Columbus, Ohio, January 28, 2000.

attached their OCLC holding symbol to a record when the URL becomes broken or redirected. A single update of the master record by one institution immediately benefits every other library.

Built using international standards including Z39.50, Unicode, Java, MARC21, and XML/RDF, the CORC system provides a scalable platform for future development.

Database

In addition to being a tool for managing electronic resources, the CORC catalog is a robust database containing more than 300,000 librarian-selected records. Using CORC, librarians select materials for their users according to their local needs. These materials cover a vast range of subject areas. The resultant database is a rich, diverse collection of materials useful to the members of the cooperative and their patrons. As more librarians use CORC and more records are contributed, the database continues to increase in scope, size and usefulness.

At the request of librarians, the CORC catalog has been synchronized with OCLC's WorldCat database. This allows for seamless integration for libraries' internal processing since many CORC users rely on the OCLC control number in WorldCat for their internal operations.

Tool for Building Pathfinders

The CORC service also provides automated tools for building pathfinders (subject bibliographies). Using CORC pathfinders, reference services librarians can share reference resources the same way that their technical services colleagues share cataloging resources. Through the use of CORC pathfinders, librarians are able to integrate their digital resources with their traditional collections. CORC pathfinders also contain dynamic searching capability that shows the user all CORC records meeting the search criteria each time the pathfinder is accessed. Pathfinders stored in CORC can automatically benefit from shared link maintenance and updates to the CORC catalog.

Technology, Cooperation, and Librarianship

Three aspects distinguish the CORC service—technology, cooperation and librarianship. CORC's technology allows for flexible integration to meet each library's needs. CORC records can be exported into the library's OPAC, or other local gateway database, or a link can be established from the library's server to pathfinders in the CORC pathfinder database. In addition, the library can use OCLC's FirstSearch 5.0, which will offer CORC records via WorldCat.

While technology provides the system infrastructure, library cooperation helps improve the service. CORC helps librarians enhance access to their local information by providing local patrons with improved access and by making their local collections available globally. By cooperatively using CORC, librarians are able to globally share their individual efforts, improve service, reduce costs, and provide their patrons with access to librarian-selected global resources.

Since becoming available to Founders' phase participating libraries in January 1999, CORC has relied on the valuable input of librarians using the service in their day-to-day activities, to help guide the design and implementation of the system. Many of the features present in the CORC system are a direct result of the suggestions and feedback provided by librarians using the service.

How Librarians are Using CORC

Librarians around the globe are using the CORC service in unique ways to manage electronic resources. The results include flexible workflow processes, distributed resource selection and description, enhanced access to information for local patrons and the ability to make valuable local collections available globally.

For example, resource selectors, collection development librarians and reference service librarians are using CORC to select electronic resources, save them in Dublin Core format, and then cataloging librarians access the records in MARC format, add authority control and further enhance the records for export to the local system.

Librarians at a museum are using CORC to describe electronic resources in Dublin Core format and then export bibliographic records in Dublin Core HTML (Hypertext Markup Language) code to their local web system for use in building web pages to be used in a database of resources.

Government librarians are using CORC to select and describe the electronic information being published by their associated agencies, to provide access to the information for constituents.

Librarians are building pathfinders that integrate into a single location, assorted resources such as books, pointers to electronic resources (CORC records and web sites) and other resources.

Some of the valuable local resources that CORC is being used to describe and make available include rare and historically significant items such as:

- Photographs of Native American Indians
- World War II posters published throughout the world
- Photographs and narratives of American Slaves
- Military speeches
- Local photographs depicting life in the past

CORC and the Digital Age

OCLC's CORC service provides librarians with the opportunity to take a leading role in the digital age. CORC combines automated tools, cooperation and librarianship to help librarians provide well-guided access to local and web-based electronic resources for their users. More information about CORC is available at <http://www.purl.org/corc>.

Automatic Web Rating: Filtering Obscene Content on the Web

K. V. Chandrinos, Ion Androutsopoulos, G. Paliouras, and C. D. Spyropoulos

Institute of Informatics and Telecommunications
National Centre for Scientific Research “Demokritos”
153 10 Ag. Paraskevi, Athens, Greece
{kostel, ionandr, paliourg, costass}@iit.demokritos.gr

Abstract. We present a method to detect automatically pornographic content on the Web. Our method combines techniques from language engineering and image analysis within a machine-learning framework. Experimental results show that it achieves nearly perfect performance on a set of hard cases.

1 Introduction

Pornography on the Internet, although less abundant than certain news reports have claimed, is a reality.¹ To cater for the broader problems of content characterization, the World Wide Web Consortium has introduced the Platform for Internet Content Selection (PICS) [6], a mechanism that allows Web pages to be rated in many dimensions (e.g. violence, nudity, suicidal content). PICS can be used either by Web authors that want to label their sites with metadata describing their content, or by third-party rating authorities. Pornographic sites are covered by at least two rating schemes, one from ICRA [7] and one from SafeSurf [8]. Once a page has been rated under PICS, popular browsers can be configured to take this rating into account. It is clear, however, that many pornographic sites are not willing to adopt self-regulation; and client-side configurations of off-the-shelf browsers can be easily circumvented.² Evidence to the fact that Internet users may elect to block pornography for themselves or minors under their responsibility comes from a number of commercial solutions. These include proxy servers that check requests against a list of forbidden or allowable URLs, and client-based software that utilizes a combination of blacklists and shallow keyword-based analysis. The dynamic nature of the Web and the fact that it is extremely easy to set up or migrate a Web site to a new IP address makes the blacklist approach ineffective.³ To counterpart this, current commercial solutions dispatch updates of their lists to their customers, with list lengths reaching up to a few hundred thousand URLs. Querying Altavista with the keyword “sex”, however,

¹ See <http://websearch.about.com/internet/websearch/library/myths> for related statistics.

² Consult [4] for an evaluation of third-party and self-regulating rating schemes.

³ An evaluation copy of a commercial product included a ~12,000 strong URL blacklist which contained only 36 out of the 500 pornographic URLs we easily summoned in a day.

returned roughly 9 million hits on average during May 2000. To alleviate list inefficiency, existing blocking software often includes keyword scanning of the URL and/or the title of the requested document. The keyword list is manually constructed to reflect frequently used terms in pornographic sites, and can often be augmented by the end-user. While real-time keyword scanning of the URL and title can improve underblocking by trapping sites not included in the blacklists, it introduces a serious amount of overblocking. For example, many Web sites maintaining content on *sexual* harassment or discrimination, *rape* prevention or even *oral* hygiene, are indiscriminately blocked under such a scheme.⁴

Our approach combines language engineering and image analysis within a machine-learning framework. It uses a probabilistic classifier trained on an appropriate corpus of Web pages, and employs both textual attributes and attributes derived from the results of an image processor. The latter estimates whether significant parts of the images on a Web page contain skin tones and are therefore suspected of depicting nude subjects. Our method is automatic in the sense that, once trained, our filter does not require Web pages to be rated manually.

2 Filtering Techniques

We represent each Web page as a vector $\vec{x} = \langle x_1, x_2, x_3, \dots, x_n \rangle$, where $x_1, \dots, x_n$ are the values of attributes $X_1, \dots, X_n$. All attributes are binary, i.e. $x_i = 1$ if the page has the property represented by X_i , and $x_i = 0$ otherwise. $X_1, \dots, X_n$ are selected from a pool of candidate attributes that includes both textual and image attributes. Textual candidate attributes correspond to words, i.e. each textual candidate attribute shows whether or not a particular word (eg. *adult*) is present on the page. There are currently only two image attributes: one showing whether or not the page contains at least one suspicious image (IMG1), and one showing whether or not the pages contains at least one non-suspicious image (IMG2). We use *mutual information* [12] to select the best attributes from the pool, and train a Naive Bayesian classifier [1] [3] to distinguish between vectors that correspond to obscene (pornographic) and non-obscene pages.

There are extensive reports in the literature on methods for skin detection, since this is a critical step towards face detection and recognition [10, 11]. These methods rely mostly on color-space transformations from RGB to $YCbCr$ or HSV where the range of skin tones is better constrained. Such transformations, however, present a trade-off between accuracy and computational expense. The only work that looks into flesh tones for the identification of naked people is that by Forsyth and Fleck [2], which attempts to use color and geometrical heuristics to detect humans. The geometrical constraints bring down recall from 79% to a mere 43%, minimizing false positive responses to 4%. Since accuracy of human bodies contouring is not as critical for our application as speed, we developed a fast and robust estimator that indicates in a single-pass whether or not more than a certain percent of an image depicts skin tones utilizing information solely from the RGB space. In our configuration, the presence of an image that has been judged to be possibly pornographic cannot alone force the

⁴ Blocked sites include Christian Bookselling Assoc. Australia and the US White House [2, 5].

classifier to block the particular page. Even a page full of images with scantily dressed people, e.g. someone's pictures from holidays on the beach, would not be classified as pornographic, since the accompanying text would not tip the classifier to that direction. On the other hand, taking images into account proved an indispensable classification aid. Although a typical pornographic Web page tends to over-advertise in text, so as to achieve a higher ranking in search-engines, there exist pornographic pages that contain very little text and many thumbnails or full-size images. These pages could not have been classified correctly without image attributes.

3 Corpus and Experimental Results

Pornographic sites are estimated to constitute around 1.5% of the Web.⁵ Attempting to maintain this analogy in the corpus would result in zero learning, because the default rule of classifying everything as non-pornographic would achieve an unbeatably high performance. Instead, we assembled a corpus that consists of pornographic pages and “near-misses”, the latter being non-pornographic pages that current blocking technology typically misclassifies as pornographic. Apart from being better for learning, a corpus of this kind pushes the filter to its limit when used for testing, as near-misses are much easier to confuse with pornographic pages than most ordinary pages. To collect near-misses, we queried a search engine with the first 10 keywords in the keyword blacklist of a widely used filtering tool. The engine was asked to return only pages that contained the keywords in their URLs or titles, which means that the pages would have been blocked by most current commercial filters. We then selected manually among the returned pages those that were not pornographic. This gave us 315 near-misses.

Collecting pornographic pages was easier, since there exist pages with many outbound links to pornographic sites. We downloaded

534 pornographic pages, a total of 849 pages. The HTML files were pre-processed to remove common words, and the remaining words were lemmatized. Ten-fold cross-validation was then performed, ranging the number of retained attributes (attributes with highest mutual information) from 5 to 250 with a step of 5. We show only “obscene precision” and “obscene recall”, as there was zero overblocking, i.e. no control page was mistakenly classified as pornographic. It appears that 100% precision can be achieved with less than 65 attributes. Recall however needs a far greater number of attributes before it can reach a comfortable 97.5%. In all cases, the

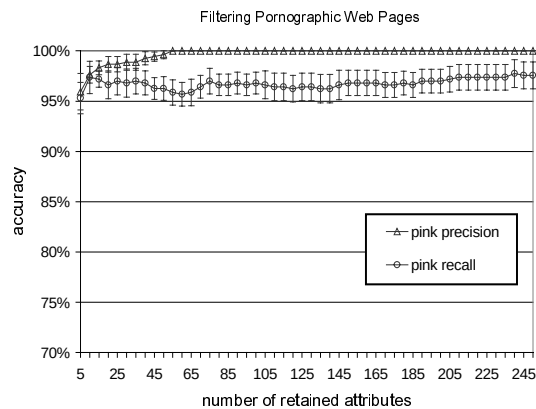


Fig. 1. Experimental results

⁵ See the study cited in footnote 1.

mutual-information attribute selection retained the IMG1 attribute (section 2), but not IMG2. We examined the few (<3%) underblocked pornographic pages to find out that they were pages with all their text in clickable images, or pages with almost no text and/or images with their color balance disrupted to achieve artistic results. We are investigating techniques to trap such pages. An initial implementation of the filtering algorithm as a proxy service accessible from standard browsers has shown very promising results, with real-time processing that maintains the scores of fig. 1 and takes a few seconds, a lag that can easily go unnoticed due to Internet latency.⁶

4 Conclusions

We presented a method for detecting automatically obscene Web content, which relies on a combination of techniques from language engineering, image processing, and machine learning. Experimental results show that our method achieves nearly perfect performance on a set of hard cases in real-time.

References

1. R.O. Duda and P.E. Hart. Bayes Decision Theory. *Pattern Classification and Scene Analysis*, pp. 10–43. John Wiley, 1973.
2. D. Forsyth. Finding Naked People. Proc. of the *4th European Conference on Computer Vision*, Cambridge, England, 1996.
3. T.M. Mitchell. Bayesian Learning. *Machine Learning*, pp.154–200. McGraw-Hill, 1997.
4. J. Weinberg. Rating the Net. 19 Hastings Comm/Ent L.J. 453, 1997.
5. Clairview Internet Sheriff, An independent review. Electronic Frontiers Australia. http://www.efa.org.au/Publish/report_isherriff.html
6. Platform for Internet Content Selection (PICS). <http://www.w3.org/PICS>
7. Internet Content Rating Association (ICRA). <http://www.icra.org>
8. SafeSurf. <http://www.safesurf.com/>
9. P. Greenfield. *Technical Aspects of Blocking Internet Content*. National Office of the Information Economy, Australia, 1999. <http://www.noie.gov.au>
10. T. Minka. *An Image Database Browser that Learns from User Interaction*. MIT Media Lab TR 365, 1995.
11. T.K. Leung, M.C. Burl, and P. Perona. Finding Faces in Cluttered Scenes Using Random Labeled Graph Matching. Proc. of the *International Conference on Computer Vision*, pp. 63–644, 1995.
12. C.D. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

⁶ See <http://www.iit.demokritos.gr/filterix> for relevant information.

The Bibliographic Management of Web Documents in Digital and Hybrid Libraries

Wallace C. Koehler, Jr.

School of Library and Information Studies, University of Oklahoma, 401 West Brooks, Room 120, Norman, OK 73019 USA
wkoehler@ou.edu

Abstract. Web documents present digital and hybrid librarians with a set of bibliographic management issues heretofore of no or minor significance for materials in print. These include frequent content change as well as the rate at which Web documents are removed or moved by their authors or creators. There have been a number of author-side and cataloger-side initiatives to assist in the management of Web documents, but these do not adequately address change. This paper explores some of those options. It addresses the impact of document change and demise on digital and hybrid collections. It offers suggestions on the management of the change and demise phenomena.

1 Introduction

The Internet dates from the early 1960s while the World Wide Web was invented in 1991. The Internet and more particularly the Web have revolutionized publication and communication. E-mail has replaced many of the functions of the telephone and the letter. Telnet and ftp permit us to manipulate and move large files. The Web has been used to replace gray literature, journals, and books. The Internet and its components must be seen in two lights. There are two sometimes conflicting definitions. The first is to define the Internet as a data transmission medium. The second is to treat it as an information source. This second definition applies primarily to the Web. Thus, not only can one employ the Web as a mechanism to access or download and publish or upload print, graphic, audio, video, or data files; one can also perceive it as the source of that information.

1.1 Digital and Hybrid Libraries

There has been an increasing interest in the development, building, and management of digital and hybrid libraries. Hybrid libraries have been defined as collections that contain both “traditional” and “digital” items [1, 2, 3]. There are a number of terms used to define libraries with digital collections. These include electronic library, virtual library, digital library, and on-line library. These terms are not, strictly-speaking, synonyms. For an elaboration of the differences and definitions, see [4]. We can, however, include all these terms under the umbrella term “digital library.”

Digital libraries are collections of organized material in digital format. Digital materials can be published, stored, and transmitted in a variety of formats. To borrow the Digital Library Foundation definition [5]:

Digital libraries are organizations that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities.

The Digital Library Foundation definition includes not only digital objects but more importantly the organization and management of those collections. Note too that it specifies “persistence over time” of the collection.

1.2 The Web as Library

The Web and the languages it supports are but one among many of the types of documents that can be maintained in a digital or hybrid library. But is the Web itself a library? I have argued elsewhere that it is not a library because it inherently lacks the organization and maintenance necessary to meet the “library” test [6].

That the Web itself is not a library does not preclude the possibility of including Web documents in digital or hybrid libraries. Indeed the inclusion of Web documents into libraries is now common practice. The many online search engines (like NorthernLight, AltaVista, and HotBot) and directories (like Yahoo!) seek to organize the Web in some meaningful fashion. The degree to which they are successful is open to question.

There are a number of more rigorous approaches to Web management. These include the recently launched OCLC Cooperative Online Resource Catalog (CORC, see <http://www.oclc.org/oclc/corc/>) project as well as a growing number projects to organize the Web. Some of these are described in Wells, Calcari, and Koplow [7]. There have been proposals and projects to archive the Web. These include Brewster Kahle’s archive (see www.archive.org) [8] as well as others as described in Hodges [9].

2 Web Management

Existing initiatives to classify and manage Web documents in libraries can be classed as either “author-based” or “cataloger-based” initiatives. Author-based initiatives are identified and applied by the document creator. Cataloger-based initiatives are value added at the library end. Some of these discussed in sections 2.1 and 2.2.

2.1 Author-Based Management

Author-based Web document management includes any mechanism that authors or Web page creators apply to their pages or sites to facilitate their identification, classification, or indexing. These range from the use of key words, spam indexing and the application of html meta tags. There are a number of more sophisticated systems. These include Dublin Core, PICS, RDF, TEI, CES, MEP, and EAD.

2.2 Cataloger-Based Management

Cataloger-based management includes those efforts that are applied after document publication. Various indexing and directory schemes, the online search engines and directories, are examples of post hoc management. While they provide an important service, that service is generally seen as inadequate for effective and quality Web document management.

There are a number of initiatives to bring effective post hoc management. These include exploitation of pre hoc or author based tools. For example, Dublin Core and other mark-up languages can be interpreted with greater or less successful by MARC based cataloging systems and incorporated into the library record. The OCLC initiatives, including InterCat, Scorpion, NetFirst, and now CORC are all attempts to apply library-based cataloging techniques to the Web. Other efforts include GILS and the proliferation of library and individual pathfinders, bookmarks, “my favorites,” jump pages, gateways, and other collections of URLs pointing to collections of “useful” and “interesting” sites. There a number of pilot sites seeking to demonstrate the utility of DDC, LCSH, or specialized thesauri as cataloging criteria.

3 Change and Demise

However valuable the pre hoc and post hoc cataloging schemes are, we must return to the problem of the changing nature of Web documents. I believe that Web documents as an information transfer mechanism lie somewhere between the ephemeral oral and unrecorded tradition and the far more permanent written record. While Web documents are very widely distributed when they are available, they are centrally owned and maintained by one or just a few individuals. Oral communication is centrally owned by the speaker, but the distribution of the message is limited. Written documents are both distributed to some number of repositories. And while ownership of the intellectual property may lie with the copyright holder, the package – including the message – is owned by someone else, oftentimes by many others [10].

Web pages and sites, as we all know, are dynamic. The decision to change, modify, move, or eliminate a Web document belongs almost exclusively to the Web author. Once a Web author elects to change a page, the changed page replaces and erases what came before. Again, as we all know, those changes may be trivial or they may be profound. This raises yet another problem. What constitutes trivial and profound changes is often subjective. Subjective differences are anathema to the cataloging process.

Graphics to support the discussions in sections 3.1 and 3.2 are provided in a slide show found online at <http://www.ou.edu/cas/slis/ethics/omega.ppt>. Space limitations preclude their presentation here.

3.1 Web Document Change

Web documents undergo change over their lifetimes. We must consider several different types of change. First, Web sites undergo modification not only to their individual pages, but to the structure of the site as well. I have found, for example, that over a period of a year, nearly every Web site will experience some change in the number of pages on the site, sometimes increasing or decreasing by several orders of magnitude [11]. In theory it is possible for Web pages to remain unchanged while the Web site of which they are a part increase or decrease in size.

Similarly, over a period of a year, almost all Web pages will also be changed [11]. There are two ways in which Web pages may be changed. First, either content may be added or deleted. Second, the number of hypertext links from the page may be increased or decreased or target page for those links may be changed. For example, if at one time a link pointed to the work of Karl Marx or V.I. Lenin but were changed to that of Groucho Marx or John Lennon, it might be argued that because the quality of the citation had changed, so had the meaning of the citing document.

I have defined changes to the page as “content” change and changes to the links from a page as “structural” change. In any given week, approximately 20% of Web pages will undergo content change at least once and again about 20% (often the same 20%) will undergo structural changes.

Some Web pages change more often than others. This change rate for individual pages may be an important cataloging element and can be captured. Once captured, that data can be populated to a bibliographic record of the page. I have termed this changed datum as “omega” [11]; the frequency over a unit time that a document changes over another unit time. I have found that the mean Web page changes about 25% of the time over one year, measured weekly.

I have also found that there is a difference in the frequency of change for navigation and content pages within a Web site as well as for pages on different top level domains (TLD). These are additional Web document elements that can be captured for cataloging purposes [12].

Navigation pages are those which direct the user to the material within the site. These are generally found within the first two levels (levels 0 and 1) of the file structure of the site; that is, the index page and those following the first forward slash (/). Subsequent pages, found following the second and subsequent forward slash usually contain the substantive information the Web author seeks to convey. Top level domains are those right most codes found on server level domains. These include the generic TLDs (gTLD) com, edu, gov, mil, org, net, and int and the country code TLDs (ccTLD). Examples of ccTLDs include .ar for Argentina, .ca for Canada, .jp for Japan, and .pt for Portugal.

3.2 Web Page Demise

Web pages are very unstable. Not only do they undergo change, they also disappear – termed comatose rather than “dead” since they sometimes return. I have monitored a sample of 360 randomly selected Web pages since December 1996. From December 1996 to April 2000, the comatose rate for the sample is about 60% gone. These Web pages disappear moreover at an almost linear rate.

Like page change, there are differences in the demise rate by level and by TLD. Navigation pages change at a rate greater than content pages, but content pages disappear at a greater rate. Over the period of study, about 70% of the content pages have gone while only 50% of the navigation pages have done so. Likewise, the ccTLDs are most likely to disappear, in part because they have redefined themselves as gTLDs. The mil, gov, org, and net pages are most likely to remain, of those only 20% have disappeared.

There are two other phenomena that should be noted. These data suggest that there is a fairly steady intermittence rate among Web pages. Intermittence is defined as those pages which “go comatose” but return within a specified period of time. That period is established here as with six collection points or five weeks. For this dataset, the intermittence rate has fluctuated between 0% and 10% with a 5% average.

Second, there has been an increase in the number of “phantom” Web pages. Phantom pages are comatose pages recognized by software as still extant. There has been a recent proliferation in the number of complex “404 Message” pages that mask the lack of content. These can be frequently recognized because there will often be a substantial decline in the size of the page (measured as bytes) which then holds constant over time. The increase in the number of phantoms has increased from nearly zero to about 12% of comatose pages.

4 Conclusions

The World Wide Web is evermore increasingly important as an information resource. Web documents are being included in digital and hybrid libraries. There have been a variety of initiatives to bibliographically manage these Web documents. These include pre hoc author-based and post hoc library based efforts.

Many Web libraries sweep their collections to weed dead links and to change links as appropriate. Unfortunately, many of these libraries do not and cannot assess existing Web pages for changes in content and in meaning.

This paper suggests that there are no easy solutions to the management of change and demise. It also suggests that the problems of change and demise are of such a magnitude that they cannot be ignored by digital or hybrid libraries seeking to maintain the integrity of their Web collections.

There are strategies that libraries with Web content can follow. These include selective selection based on the likelihood of persistence, frequent sweeping and measurement of document sizes and structures, and periodic content or meaning assessments. The “behavior” of Web pages can be incorporated into the bibliographic record – the “card catalog” or OPAC – to alert the end user to the probability of change any given document may undergo over time.

References

1. Pinfield, S., Eaton, J., Edwards, C., Russell, R., Wissenburg, A.: Realizing the Hybrid Library. D-Lib Magazine (1998)
<http://mirrored.ukoln.ac.uk/lis-journals/dlib/dlib/october98/10pinfield.html>
2. Rusbridge, C.: Towards the Hybrid Library" D-Lib Magazine (1998)
<http://www.dlib.org/dlib/july98/rusbridge/07rusbridge.html>
3. Lesk, M.: Practical Digital Libraries: Books, Bytes, and Bucks. Morgan, Kaufman San Francisco (1997).
4. Koehler, W. (2000)
<http://www.ou.edu/cas/slis/courses/LIS5990A/slis5990/DigLib/index.htm>
5. Digital Library Foundation, <http://www.clir.org/diglib/dldefinition.htm>
6. Koehler, W.: Digital Libraries and World Wide Web Sites and Page Persistence. Information Research (1999) <http://www.shed.ac.uk/~is/publications/infres/paper60.html>
7. Well, A.T., Calcari, S., Koplow, T.: The Amazing Internet Challenge: How Leading Projects Use Library Skills to Organize the Web. American Library Association Chicago (1999)
8. Kahle, B.: Preserving the Internet. Scientific American (1997)
<http://www.sciam.com/0397issue/0397kahle.html>
9. Hodge, G.M.: Best Practices for Digital Archiving: An Information Life Cycle Approach. D-Lib Magazine (2000) <http://www.dlib.org/dlib/january00/01hodge.html>
10. Koehler, W.: The World Wide Web as a Third Information Model: Revolution or Old Wine in New Bottles? Crimea 98, Libraries and Associations in the Transient World: New Technologies and New Forms of Cooperation Proceedings (1998)
<http://www.gpntb.ru/win/inter-events/crimea98/doc1/doc65.html>
11. Koehler, W.: An Analysis of Web Page and Web Site Constancy and Permanence. JASIS (1999) 162-180.
12. Koehler, W.: Classifying Websites and Webpages: The Use of Metrics and URL Characteristics as Markers. Journal of Librarianship and Information Studies (1999) 21-31.

The Economic Impact of an Electronic Journal Collection on an Academic Library

Carol Hansen Montgomery, Ph.D.¹ and John A. Bielec, Ph.D.²

¹ Drexel University, W. W. Hagerty Library, 3141 Chestnut Street, Philadelphia, PA 19104, USA
<http://www.library.drexel.edu/facts/staff/dean.html>
montgoch@drexel.edu

² Drexel University, Office of Information Resources and Technology, 3141 Chestnut Street, Philadelphia, PA 19104, USA
<http://www.pages.drexel.edu/~jbielec>
jbielec@drexel.edu

Abstract. This paper provides information on the economic impact of the transition from print to electronic journals in an academic library. The technological orientation of the university and a robust network infrastructure have made it possible for Drexel to make this transition more quickly than most, if not all, U.S. academic libraries. Shifts in costs occur in all budget areas: capital (space and network infrastructure), staffing, purchased services, materials, supplies and equipment. Overall, costs are higher, but preliminary data indicates that in Drexel's case per journal and per article costs are lower.

1 Introduction

In 1998, the W.W. Hagerty Library of Drexel University made migration to an electronic journal collection as quickly as possible a key component of its strategic plan. If a journal is available electronically, **only** the electronic version is purchased whenever possible. The sole exceptions are (1) when the electronic journal lacks an important feature of the print version (e.g., advertisements in business or ads in fashion journals) and (2) when the journal is part of the browsing collection (e.g., *Scientific American* and *Harvard Business Review*). By the summer of 2000, Drexel's journal collection consisted of 950 print subscriptions and 6000 electronic journals. A dramatic change in staff workload is the most immediate impact on library operations, but space, equipment, and even supply needs are affected.

2 Development Stage

Drexel's largest college is Engineering. The library's "fast track" transition to electronic journals was made possible by the Drexel environment which includes not

only a supportive administration and technologically-savvy faculty and students, but an advanced computer network.

2.1 Wireless Technology Provides Access Anywhere

Drexel University made history when in 1983 it required all students to have access to a personal computer. Drexel quickly became known as one of the nation's foremost technological universities. This lead continued with one of the first ResNet (fully wired student residential facilities) in the nation including Ethernet, and full participation in the Next Generation Internet initiatives and Internet2 consortium. In early 1998 Drexel University's first "CyberZone" was completed through a cooperative effort between the Office of Information Resources and Technology and the Hagerty Library. Students, faculty and staff were able to sign-out laptop computers and access via the latest wireless technology any system on the Drexel network including email, the library catalog and digital resources as well as the World Wide Web. This created a whole new era of opportunity for Drexel students wanting to combine various electronic and database information sources with traditional library resources independent of location, and became the foundation for many library improvements, including a major shift from the print to digital format. The CyberZone concept provided access to digital library resources from literally any place at any time. Today, CyberZones span the entire campus, creating a CyberCampus that makes ubiquitous computing at Drexel a reality. As a result of these technological initiatives, Drexel University is ranked as Philadelphia's top wired university and the 16th most wired college in the United States by *Yahoo! Internet Life* magazine in its annual survey of "America's 100 Most Wired Colleges."

2.2 Development of Drexel's Electronic Journal Collection

In concert with development of the robust, state of the art, electronic infrastructure, library staff began developing and refining selection methods for electronic journals. The selection/ordering/acquiring process is far more complex than that in place for print journals. The current and evolving pricing models and methods for purchase often include buying packages of titles or services, many with value-added features. Reviewing and negotiating proper terms for e-journal licenses is a major aspect of the complexity. Additionally, new variables must be considered in the title-by-title evaluation of available content (e.g., graphics, linking options, web interface functionality, and other value-added features).

Whenever possible, we purchase only the electronic version of a journal and cancel the corresponding print publication. This has met with limited success so far. One of our most interesting problems has to do with creative methods NOT to receive print journal issues when they are bundled with the electronic version. Because of the added cost of receiving, processing, binding, shelving and storing the print format, staff developed several strategies for eliminating print issues when the library has purchased equivalent electronic access.

3 Impact on Library Staffing and Other Costs

Table 1 delineates many of the changes in staffing and other costs. See [1] for more detail.

Table 1. Changes in staffing and other costs

Department	Activity	Net Impact
Infrastructure/ Systems&Space	☐ campus network	↑ increased capital costs
	☐ computer hardware	↑ increased equipment costs
	☐ computer systems maintenance	↑ increased staffing
	☐ hardware maintenance	↑ increased costs
	☐ setting up access	↑ increased staffing
	☐ software purchase & development	↑ increased staffing
	☐ printing	↑ increased costs & revenue
	☐ space utilization	↓ reduced space needs
Administration/ Management	☐ negotiating contracts	↑ increased staffing
	☐ managing the change	↑ increased staffing
	☐ attention to decisions	↑ increased staffing
	☐ budgeting	↑ increased staffing
	☐ subscription fees	↑ increased costs
Technical Services	☐ print journal check-in	↓ reduced staffing
	☐ acquisitions	↑ increased staffing
	☐ claiming	↓ reduced staffing
	☐ bindery staffing effort and fees	↓ reduced staffing & costs
	☐ cataloging new items	↑ increased staffing
	☐ OCLC transactions	↑ increased costs
	☐ catalog/e-journal list maintenance	↑ increased staffing
Circulation/ Access	☐ re-shelving journals	↓ reduced staffing
	☐ collecting use data	? net impact unclear
	☐ stack maintenance	↓ reduced staffing
	☐ user photocopying	↓ reduced use & revenue
Reserve	☐ article file maintenance	↓ reduced staffing
	☐ article checkout	↓ reduced staffing
	☐ maintaining e-reserves	↑ increased staffing
Document Delivery	☐ faculty copy service	? net impact unclear (started 1999, expect reduction)
	☐ interlibrary loan - borrowing	↓ reduced staff costs
	☐ interlibrary loan - borrowing	↓ reduced vendor charges
	☐ net ILL volume	= expected to decline
Information Services	☐ reference at desk	? net impact unclear
	☐ instruction/promotion	= expect increase
	☐ preparing documentation	↑ increased staffing
	☐ journal selection	↑ increased staffing

In summary, the impact on overall costs and staffing patterns has been pervasive and substantial. Preliminary and partial analysis of the staff cost changes alone indicate that capital investment, i.e. system and other development costs allocated to e-journals, is over \$100,000, and that the annual operational increase is at least \$80,000. See [1] for detailed changes by category.

The primary reason for the increase is that the activities associated with electronic journal acquisitions and maintenance are far more complex, and thereby more time-consuming, than those required for print journal collections. And, because of the complexity and skills required by the use of technology, the staff required to manage e-journals demand higher salaries than staff devoted to the print journal collection. There are real and measurable savings in the print area; for example, in journal check-in, binding, claiming, repairing, shelving and statistics gathering take less time. But these are routine tasks typically performed by the lowest paid clerical staff and student workers. But these savings are more than offset by the activities (e.g. setting up access via URLs, fixing broken links) required to maintain an e-journal collection.

Journal selection in the print environment is not simple but has many fewer dimensions than the electronic analog. In the e-journal world many more factors come into play when making collection development decisions. Licensing restrictions, quality of visuals, inclusiveness of all the features in the print counterpart, whether the journal is part of an aggregator's list, overlap among aggregator's listings, presence of a login requirement, interface, and linking capabilities are examples of added characteristics that must be considered. "Deals" are negotiated with publishers and aggregators by local consortia; alternate sources must be compared to obtain the best pricing. Most acquisitions involve contracts. Some vendors and publishers have very complex pricing plans. All this evaluation and weighing of alternatives must be done by senior professional staff. This process, plus managing all the personnel and budget shifts, and other changes [2], places a large burden on library administrators.

Considering only subscription costs, Drexel's current e-journal collection is far less costly than the print collection on a per title basis. Obviously, this is the result of the composition of the two specific collections that are not comparable and so is not generalizable. Yet the difference is so striking that it is worth mentioning. The per title cost of the Drexel e-journal collection is less than half that of the print collection, a figure even more notable if one considers that e-journal access almost always includes several back years. [1] Finally, the use of our e-journals is double that of print and rising rapidly, so the savings for per article use is even greater. [1]

References

1. Montgomery, C., Sparks, J. Framework for Assessing the Impact of an Electronic Journal Collection on Library Costs and Staffing Patterns. Mar. 2000.
<<http://www.si.umich.edu/PEAK-2000/montgomery.pdf>> (25 June 2000).
2. Montgomery, C., and Sparks, J. The Transition to an Electronic Journal Collection: Managing the Organizational Change. *Serials Review*. (In press)

A Comparative Transaction Log Analysis of Two Computing Collections

Malika Mahoui and Sally Jo Cunningham

Department of Computer Science, University of Waikato, Hamilton, New Zealand

Telephone: +64 7 838 4021 Fax: +64 7 838 4155

Email: {mmahoui, sallyjo} @cs.waikato.ac.nz

Abstract. Transaction logs are invaluable sources of fine-grained information about users' search behavior. This paper compares the searching behavior of users across two WWW-accessible digital libraries: the New Zealand Digital Library's Computer Science Technical Reports collection (CSTR), and the Karlsruhe Computer Science Bibliographies (CSBIB) collection. Since the two collections are designed to support the same type of users—researchers/students in computer science—a comparative log analysis is likely to uncover common searching preferences for that user group. The two collections differ in their content, however; the CSTR indexes a full text collection, while the CSBIB is primarily a bibliographic database. Differences in searching behavior between the two systems may indicate the effect of differing search facilities and content type.

1 Introduction

Transaction log analysis has been extensively applied to the study of searching behavior in OPACs (for an overview of this body of literature, see [3]). Log analysis has the advantage of providing a descriptive summary of large numbers of user interactions, and as such is a valuable research tool for investigating user search behavior. As a caveat, it is difficult to generalize for all users, at all times—the value of the technique lies in describing a particular system's users, searching specific collections. The contribution of this paper is to provide an opportunity for generalization by examining the logs for two digital libraries created to support the same type of user: researchers and tertiary students in computer science.

Digital libraries usage logs have only recently attained a usage level appropriate for log analysis (see, for example [1, 4]). In a previous paper [2], we applied transaction log analysis techniques to the Computer Science Technical Reports (CSTR) collection of the New Zealand Digital Library (<http://www.nzdl.org>). At that time, the CSTR indexed nearly 46,000 publicly available computing-related technical reports harvested from over 300 research institutions from around the world. As no formal bibliographic metadata is available for the CSTR documents, the search interface is based solely on keyword searching, and the full text of the documents have been indexed. When searching identifies a relevant

document, the document can be immediately retrieved; although the CSTR itself does not store copies of the collection documents, the collection provides links to the original sites from which the documents were identified for indexing.

In contrast, the Computer Science Bibliographies (CSBIB) collection primarily contains bibliographic metadata: the collection holds over a million references in BIBTEX format, donated as single references, subject bibliographies, or single-author bibliographies (<http://liinwww.ira.uka.de/bibliography/index.html>). Over 90,000 of the references contain a link to an online version of the corresponding paper; however, the full text of the papers is not searchable, and less than 1% of the references include links to papers. The emphasis of the CSBIB collection is the provision of references, rather than immediate access to the documents themselves.

In the following section we describe how the transaction log data has been collected. The usage logs are automatically processed for analysis. The results of this analysis are presented in Sections 3 - 5: we describe query complexity and query types, analyze user session lengths, and examine user acceptance of system default settings. These statistics are compared across the two collections, and (where relevant) across different search options within each system.

2 Data collection

All user activity in the CSBIB and CSTR collections is automatically logged. Although actions can be grouped into user sessions, individual users cannot be identified—the users of the two digital libraries remain anonymous. For the CSTR collection, local users were excluded from this analysis; during the period studied many local queries were submitted as system tests.

The CSBIB log totals 251878 queries submitted in the period 1st September 1999 to 26th December 1999. CSBIB queries are of two types: simple search (keyword search over all bibliographic record fields) and advanced search (in which users specify which fields to search). For the CSTR, over 30000 queries were recorded from April 1996 to July 1997. User activities in both logs are timestamped. The log attributes used in this analysis include the query text and query options (ie, simple or advanced for CSBIB, ranked or Boolean for CSTR).

3 Query complexity

Queries to both collections are short; over 90% of the CSBIB simple search queries and approximately 80% of CSTR queries contain three or fewer terms (Table 1). The average number of search terms in a CSBIB simple search query is 1.8%, compared to 2.5% in the CSTR collection. This difference is slightly misleading, in that the CSBIB syntax permits author names to be entered with initials appended to the family name (for example, as SmithJ rather than Smith J or John Smith)—which will have the effect of reducing the number of query terms in many author queries.

Table 1. Distribution of the number of query terms (CSBIB simple search, CSTR)

No of terms in query	0	1	2	3	4	5	6	>7
CSBIB (simple search)	0%	52.72%	28.10%	10.8%	4.18%	1.75%	1.02%	1.41%
CSTR	1.59%	27.06%	34.04%	19.76%	8.98%	4.26%	2.06%	2.25%

In both systems the default Boolean operator is an OR, and the majority of queries contain no explicit Boolean operator. The most common user-specified Boolean operator is the AND (Table 2). Interestingly, in the CSBIB collection

Table 2. Frequency of operators in Boolean queries.

percentage of queries containing	CSTR	CSBIB
no Boolean operators	66.0%	84.1%
at least one intersection operator	25.8%	14.18%
at least one union operator	2.5%	1.69%
parentheses for compound expressions	4.6%	0.01%

only 14% of the queries included the AND operator compared to 26% in the CSTR collection—despite help text on the CSBIB simple search page that concisely enumerates the logic operators and gives a small example on how to use them. The CSTR interface does not contain syntax help on the search page itself.

4 User sessions

Each query in the transaction logs contains a user id (although, as noted in Section 2, these ids cannot be traced back to collection users). A simple heuristic was used to identify user 'sessions': a session is assumed to be a series of queries containing the same user id, and with no more than a 30 minutes lapse between consecutive queries.

For both the CSTR and the CSBIB collection, users tend to issue relatively few queries per session (Table 3); over 80% of user sessions include 5 or fewer queries. The interesting difference between the CSTR and CSBIB collections is the relatively larger number of CSBIB sessions that include more than 9 queries. It appears that CSBIB users tend to persevere in query refinement to a greater extent than CSTR users. We can only hypothesize about the basis for this difference. It may indicate that the CSBIB interface encourages a greater degree of exploration; or that the users of the CSTR tend to require more exhaustively complete results (a plausible hypothesis, if CSBIB users are more likely to be students/academics seeking to complete literature reviews); or that the ability to easily view the full text of search results (as in the CSTR) permits a user to more quickly home in on relevant documents in the collection; or perhaps

Table 3. Frequency distribution of the number of queries issued in user sessions.

No queries issued in a user session	CSTR % of sessions	CSBIB (simple search) % of sessions	CSBIB (advanced search) % of sessions
1	43.89	35.97	29.95
2	21.95	20.02	20.43
3	12.1	12.19	12.88
4	7.76	8.51	8.46
5	4.88	5.84	5.82
6	2.90	3.83	4.22
7	1.92	2.68	3.14
8	1.53	2.13	2.35
>9	2.41	7.32	10.79

there are other explanations. This point indicates a weakness of transaction log analysis: it can indicate patterns of user behaviors, but cannot explain those behaviors. At this point, we must engage in an interview-based or ethnographic study to further explore the users' motivations.

Tracing the individual user ids across the time periods captured in the transaction logs, we can determine the number of repeat visitors to the CSTR and CSBIB digital libraries (Table 4).

Table 4. Distribution of repeat visits to the CSTR and CSBIB collections.

number of visits	CSTR (%)	CSBIB (%)
1	72.82	72.48
2	14.36	11.24
3	4.31	4.24
4	2.19	2.99
5	1.42	1.65
6	1.06	1.49
> 6	3.84	5.88

Again, the two logs indicate similar behavior. Disappointingly, nearly three-quarters of the users of both collections visited the collection only once during the extensive time period covered in this analysis. The result is perhaps tied to the relatively large proportion of users who issue only one or two queries during a search session—and who presumably either have very straightforward information needs that are quickly satisfied, or who decide that the collection cannot fulfil their information needs.

Analysis of consecutive queries in a session reveals an interesting difference in query refinement behaviour: more than half (58.38%) of the queries in CSBIB

sessions have no term in common with the previous query, compared to only 33% of the consecutive queries in the CSTR collection (see Table 5).

Table 5. Frequency with which consecutive queries contain common terms (CSBIB simple search, CSTR).

No of terms in consecutive queries	0	1	2	3	4	5	>5
CSBIB (simple search)	58.38%	25.85%	10.70%	2.98%	1.11%	0.45%	0.51%
CSTR	33.53%	22.56%	23.08%	11.34%	4.71%	2.22%	2.25%

These figures discount the first query in a session. This low incidence of term overlap in CSBIB collection indicates either that CSBIB users tend to attempt to satisfy more than one distinct information need in a session, or that the CSBIB users refine consecutive queries more radically than CSTR searchers.

5 User acceptance of default settings

The logs from both collections show that users tend to settle for the system's default settings. For the CSTR, during the log collection period the default search type was changed from Boolean to Ranked—and approximately 66% of queries used the default setting, no matter what the setting actually was (Table 6). Approximately 80% of queries to the CSBIB collection were submitted through

Table 6. Frequency of default query types.

CSTR	Boolean as default 46 week period	Ranked as default 15 week period	CSBIB	
number of queries	24687	8115	simple queries	202947 (80.57%)
Boolean queries	16333 (66.2%)	2693 (33.2%)	advanced queries	48931 (19.43%)
Ranked queries	8354 (33.8%)	5420 (66.8%)		

the default standard search (keyword search) interface. Given the strong tendency of this user group to accept system defaults, it is important that these defaults be set 'correctly'—that is, so as to maximize the opportunities of the users in satisfying their information needs. Given that users tend to submit short, simple queries (Section 3), a useful interface strategy may be to maximize search recall for those queries. To that end, the choice of ranked output as default for the CSTR and simple search (keyword, over all bibliographic record fields) for the CSBIB appear to be sensible interface design decisions.

6 Conclusions

This paper updates a previous study on transaction log analysis ([2]) by comparing the searching behavior of users across CSTR and CSBIB, two WWW-accessible digital libraries for computer science researchers. Commonalities in the log analysis results indicates that this user group prefers to issue relatively few, brief queries in a session. While computing researchers might be expected to actively explore software and its functions, these users tend to accept default settings—no matter what those settings are. Differences in search behavior across the two systems include a tendency of CSBIB users to issue more queries in a session, and for consecutive queries in a session to show a lesser degree of term overlap. Further study (including qualitative, interview-based examinations of small groups of users) is indicated to determine the basis for these differences.

Acknowledgements

Our thanks go to Stefan Boddie for his contribution in the preparation of the data. Alf-Christian Achilles has developed and maintained the CSBIB collection since 1995; he generously provided the CSBIB transaction logs analyzed in this paper.

References

1. Jansen, B.J., Spink, A., Saracevic, T.: Failure analysis in query construction: data and analysis from a large sample of web queries. *Proceedings of ACM Digital Libraries*, Pittsburgh (1998) 289-290
2. Jones, S., Cunningham, S.J. and McNab, R.: An analysis of usage of a digital library. *European Conference on Digital Libraries*, Heraklion, Crete, Greece, August. *Lecture Notes in Computer Science* **1513** (1998) 261-277
3. Peters, T.A.: The history and development of transaction log analysis. *Library Hi Tech* **42**(11:2) (1993) 41-66
4. Spink, A., Batement, J., and Jansen, B.J.: Searching heterogeneous collections on the web: behaviour of Excite users. *Information Research: an electronic journal* (1998) 4:2

ERAM - Digitisation of Classical Mathematical Publications

Hans Becker and Bernd Wegner

SUB Göttingen and TU Berlin, Germany,
becker@mail.sub.uni-goettingen.de, wegner@math.tu-berlin.de

1 Synopsis

Longevity is typical for research achievements in mathematics. Hence to improve the availability of the classical publications in that area and to enable to get quick information on these, electronic literature information services and digital archives of the complete texts will be needed as important tools for the mathematical research in the future. This will bring the holdings from the journals archives nearer to the user and will prevent lost of the papers because of the deterioration of the paper as a consequence of age.

The aim of this article is to give a more detailed report of a project, the so-called ERAM project, capturing the "Jahrbuch" as a classical bibliographic service in mathematics in a database and using this activity to select important publications from the Jahrbuch period for digitisation and storage in a digital archive. The database will not be just a copy of the printed bibliography. It will contain a lot of enhancements like modern subject classifications as far as possible, keywords giving ideas about the content in modern terms and comments relating classical results to modern mathematical research areas. These features will remain open for additions within a living project. The digital archive built up in connection with the database will be linked to the database and provide all facilities associated with current digitisation projects. The content will be distributed to mirrors and combined with similar archiving activities in mathematics.

2 Introduction

In order to understand the importance of the availability of classical mathematical publications for current research in mathematics some remarks concerning the interest of mathematicians in these publications should be made at the beginning of this article. While generally in science the interest in the author's personal views of a topic is the main motivation to consult a publication, mathematicians are looking for precise results like statements and proofs of theorems. Any reliable source to find these theorems will be sufficient for them. But a lot of these results are only stated at one place, the paper where they have been published first. Even if these results have not been reproduced in monographs or

surveys later on, this does not indicate that they became obsolete. Mathematical knowledge does not age.

In addition to these results special aspects of their proofs may be of later interest. This requires to have easy access to these documents, no matter what their age and state are. Searchability will be an additional requirement to enable the researcher to find his way in the huge knowledge base of mathematical achievements. Admittedly, no current search engine is able to locate a statement in its abstract meaning. Names for some of them will help, and classification codes of special subject areas will restrict the set of documents where to look for the desired information considerably. Hence literature databases for the classical period of mathematics are desirable. They should offer the same facilities like the current literature information services in mathematics, and even more, they should also provide links to the future given by modern mathematics. This is the starting point for the project ERAM which also will be called the Jahrbuch-project for short.

The acronym ERAM stands for "Electronic Research Archive for Mathematics". The project is funded by the Deutsche Forschungsgemeinschaft (DFG). The institutions caring about the project are the Staats- und Universitätsbibliothek Göttingen (SUB) and the Technische Universität Berlin (TUB). Being the scientifically more challenging part of the project, main attention is paid to the database in this article. Nevertheless, some hard work has to be invested into the installation of the digital archive. But the unpredictable part of this business are the negotiations with publishers to get the licenses for offering the content of the archive at a low rate only, which may cover the costs for the maintenance of the archive, not those of the installation.

3 Structure and Enhancements of the JFM-Database

The Jahrbuch über die Fortschritte der Mathematik (JFM) was founded in 1868 by the mathematicians Carl Ohrtmann and Felix Müller. It appeared in 68 issues from 1868 - 1943. Commonly one issue contained all the reports of mathematical papers which appeared in the year mentioned on the issue, but some issues contained more than one year. More than 300.000 mathematical publications in the above period were reviewed by the JFM. Thinking of the needs of ERAM the JFM is a perfect source to build up the electronic gateway to the digital archive.

The first step of the ERAM-project is the production of a bibliographic database, the JFM-database, capturing the content of the Jahrbuch über die Fortschritte der Mathematik (JFM). As a matter of principle the JFM-database should be accessible with the same search software like the well-established database for mathematics, Zentralblatt MATH. Hence it should provide a similar structure of fieldings like Zentralblatt and it should implement the same bibliographic standards. But it will not suffice just to bring the content of the JFM in an electronic form. Modern literature databases provide several search options for which the information could not easily be extracted from the text of the JFM.

Editorial enhancements will be needed, and moreover historical links should be provide to modern research as far as possible.

For example, the only formalised subject information in the JFM consists of the subject headings which are stored in the database like a raw classification. A more precise description of subjects could only obtained by additional intellectual work relying on the support of mathematical experts. Standardisation and links to articles involve the additional editing of the data by librarians. In addition to this there will be an evaluation of the publications providing some ranking according to their importance. This ranking can be a superficial one only. Its only purpose is to decide on the relevance of the distinct documents for the digitisation of the full text.

More detailed updated information for the papers will be found under "expert's remarks". The expert may use this field for an annotation to the publication of any kind. This could be a verbal description of the importance of the publication or the reference to other developments in mathematics, for example. We hope that in particular those classical papers which initiated a whole series of research papers or even a new subject area in mathematics can get a special representation in the JFM database. Going back and forth with links that may enable the user to find relations between current publications having the same historical root without being part of the same modern mathematical subject. The content of this field will depend on the knowledge of the expert handling the corresponding document, and in general such remarks initially only will be found with a few papers in the database.

Another set of enhancements of the JFM-database will be the result of the editing of data by librarians. They will care about the standardisation of the information available and the provision of links to digital versions of the corresponding article or a library where the article can be ordered using a document delivery service. For example, the reviews in JFM cite sometimes former publications which have received an identifier within the JFM-database. These references will be transformed in executable hyperlinks. URLs of a digitised form of the original publication are specified, and the signatures of the source of the original publication at the SUB are added. SUB provides a document delivery service for these publications.

4 The Digital Archive

In addition to its usage as high-quality source for information on classical mathematics, the JFM-database will provide access to the digital archive to be built up within the project. This archive should contain most of the relevant mathematical publications from the period of the JFM (1868 - 1943). A principal rough estimate of the amount of documents which could be covered by ERAM gave a rate of about 20

The content of the archive will be stored at SUB and made accessible there. Co-operation with other digital archives like mirroring of content and exchange of documents will be a matter of future discussions. A search in the JFM-database

is the easiest way for accessing the archive. With the hit list of his search the user will get hypertext links to the document, if it is available in the digital archive, or to the order form of the document delivery service, if the document is available in the SUB-holdings at all.

The selection of the documents is based on the recommendations from the JFM-experts in principle. But this turned out to become a bottle neck of the project. Moreover, in several cases scanning distributed patches from journals and other sources is less efficient than scanning an issue of a journal as a whole. This lead to the conclusion that additional selection methods should be applied, without ignoring the advice from the experts at all. In addition to this the restriction of the selection to the period of the JFM is rather artificial, though this had been an important range to start with. Hence the supervisors of the project started to add recommendations for documents to be scanned on their own.

According to hints from the users of a first test version of the database, the access even to the raw data seems to be an important facility for the mathematical community. Hence everything which has been put into the system is made freely accessible under the URL <http://www.emis.de/projects/> clicking on the box for the Jahrbuch. Users are alerted that the content is in a preliminary state only in many cases. About 60

The accessibility to first parts of the archive is under preparation. This partially waited for having a critical set of contents in the JFM-database. But also the structures for the public access had to be discussed and installed. Hence the archive has more documents available than can be seen from outside, and the content is growing permanently. Also discussions have been initiated to link the JFM-database with other digital offers of documents from the JFM-period and to exchange documents with them. Hopefully this will lead soon to a distributed system of digital archives for mathematics, for which the JFM-database can be used as the most convenient gateway.

The Electronic Library in EMIS - European Mathematical Information Service

Bernd Wegner

TU Berlin, Germany
wegner@math.tu-berlin.de

1 Preliminaries

The idea to develop the European Mathematical Information Service EMIS was born at the meeting of the executive committee of the EMS (European Mathematical Society) in Cortona/Italy, October 1994. It was decided to set up a system of electronic servers in Europe for Mathematics under the auspices of the EMS, and this was extended very soon to the current version of a central server collecting mathematical information and distributing this through a world-wide system of mirror servers. The installation of the central server began in March 1995 in co-operation with FIZ Karlsruhe at the editorial office of Zentralblatt für Mathematik in Berlin. In June 1995 EMIS went on-line under the URL <http://www.emis.de/>. The first mirrors were established very soon in Lisbon, Southampton and Marseilles.

The main sections of the contents of EMIS are the Electronic Library, information about the EMS, the European Mathematical Congresses, a link to the database MATH of Zentralblatt für Mathematik, an access to directories of mathematicians, and information about other mathematical servers throughout the world. World Wide Web access is regarded the primary access method to EMIS. The access to the contents of EMIS is free, and also the link to the database contains a free component which will be explained later in the corresponding section of this article.

The most important general aspect of EMIS is the idea to distribute the service through a world-wide system of mirrors where the full content of the service will be available and updated periodically. This improves the accessibility of EMIS, and it simultaneously is important for the safety of the data and their archiving: if the master server one component of the system crashes down, it can be regenerated easily from the other components. In principle, every European country should have one mirror at least, and two mirrors should not be too near to each other in terms of the network geometry.

2 The Electronic Library

The Electronic Library of EMIS has the aim to present a collection of freely accessible electronic publications which should be as comprehensive as possible. There are three sections: electronic journals, electronic proceedings volumes, and

electronic monographs. In order to guarantee that the electronic publications stored in the Electronic Library meet the requirements satisfied by articles in print journals, the decision on the inclusion of journals, proceedings or monographs is taken in accordance with the Electronic Publishing Committee of the European Mathematical Society. Hence no items will enter the library which have not been evaluated and recommended by a referee within the editorial procedures of the corresponding series. This is in particular important in order to rule out the reservations of many mathematicians who have the opinion that electronic publishing will damage the quality of mathematical publications.

The section on Electronic Journals contains pure electronic journals as well as electronic versions of print journals. The pure electronic journals are produced elsewhere and EMIS only serves as an additional distributor. The installation of electronic versions of print journals depends on the technical facilities of the editors of these journals. Some of them prefer to offer the electronic version on their home server, such that a mirror of this content can be taken over by EMIS. Others do not care about the installation of the electronic version themselves, and send the files to the master site of EMIS in Berlin where they are brought into a shape which is suitable for an electronic offer through WWW. Most of these print journals are published at a low-budget level, and hence the risk to loose subscribers to the print version due to the free electronic offer is considered as low by them.

Some titles stored in the electronic library are as follows where the pure electronic journals are marked by an (e) in this list of samples:

Annales Academiae Scientiarum Fennicae Series A. Mathematica (Helsinki),
 Beiträge zur Algebra und Geometrie / Contributions to Algebra and Geometry (Berlin),
 The Electronic Journal of Combinatorics (e),
 Electronic Research Announcements of the AMS (e),
 Electronic Transactions on Numerical Analysis (e),
 Journal de Théorie des Nombres de Bordeaux (Bordeaux),
 Matematicki Vesnik (Belgrade),
 Mathematical Physics Electronic Journal (e),
 Portugaliae Mathematica (Lisbon),
 Revista Colombiana de Matemáticas (Bogotá), et al.

The access to these journals in EMIS is organised quite conventionally. On the home page of EMIS a list of mirrors is provided where the site with the (probably) quickest access can be clicked. Then a choice can be made, to enter the Electronic library through the short list of journals without graphics or to use the full display of these items. The full display contains also background information on the editorial policy of the corresponding journal and instructions how to submit an article. Having made the choice of journal one wants to read, the level of the contents will be reached which is organised as usual. At that level information on the offer of files is given. In all cases DVI- and Postscript-files will be available, sometimes also TEX-files can be found in addition to that, and in most cases PDF files will be offered. Also printing or storage of these files will be possible at the site of the user, but he is requested to

respect the copyright according to the rules of the corresponding journal. The access to the section of Proceedings Volumes is organised in a similar way.

EMIS with its mirrors is just the appropriate system to improve the distribution accessibility of electronic publications. This is underlined by the present access numbers which are growing continuously. The additional offer of electronic versions by the system of mirrors of EMIS is used frequently: all servers together report some 200,000 hits per week.

3 The Links to the Databases

To connect EMIS with the databases of Zentralblatt MATH and MATHDI is one part of the increasing involvement of the European Mathematical Society in the edition of this reviewing service. The vision for the near future is to design this service as a European database in mathematics, which has Zentralblatt as its core, but relies on additional distributed input from several European backbones. Hence this information service should continue the tradition of Zentralblatt to cover the mathematical literature as completely as possible by transferring some part of the responsibility to these backbones and giving them the opportunity to take better care about the representation of the achievements of their mathematical community. This is the goal of the project LIMES funded by the 5. Framework Programme of the EU.

EMIS provides a link to the WWW-access which can be reached directly under the URL <http://www.emis.de/ZMATH>. Additional access will be directly available to the mirrors of MATH at seven other sites world-wide. MATH covers information on all mathematical literature, starting with 1931. The total amount of documents for which information is stored in MATH is more than 1.700.000. This increases by more than 65.000 items annually. The update of the database is made every two weeks which corresponds to the production of the print version of Zentralblatt. The scope of publications handled by this service includes in addition to mathematics mathematical applications in physics, mechanics, computer science, economics and biology.

Searches can be made using the following list of fields: authors, titles, classification, basic index, source, language etc. A search can be formulated in logical combinations of these terms. The search can be made in a command mode or be guided by a choice of graphic menus. The information is available in the AMSTEX source code, but tools for a convenient formula display are available. Download of the hit list at the users site is possible, and, coming back to the other content of EMIS and other electronic journals and publications, links to the full text of the corresponding article will be arranged, if this is available electronically. One additional option to get the full text of such an article consist of central document delivery services. Buttons to connect to such services and to see, if the corresponding article is available there, are installed in the menu for the search. Document delivery can be arranged by these services electronically or by sending copies by ordinary mail at reasonable rates.

Finally, it should be mentioned that there is a free component in the link from EMIS to MATH. Any user of EMIS can do searches in MATH. But non-subscribers only will get information on at most three hits from their list. These hits will be taken from the top of the list, where the most recent items are listed. Hence, if very precise information is searched which probably leads to a small hit list, then this can be reached by everyone who has access to the internet.

4 Future Developments

In combination with the Electronic Library and other links to electronic publications the offers of MATH and other databases in EMIS provide a quick and easy access to qualified mathematical publications. These tools can be used by the mathematical community at suitable sites simultaneously for free or at modest rates. But the system will have to care about the implementation of new facilities which came up with the general development of electronic libraries.

EMIS will have to handle the journals which provide printed and electronic versions simultaneously in a different way. There should be a change to "online first". This will speed up the publication of the papers and it will give considerable advantage to the electronic version. Presumably there will be no chance that these journals can survive from the income for their print version only. Furthermore PDF turned out to be the most popular format for reading electronic journals, also in mathematics. Hence the offers in the library of EMIS should provide PDF-files in the near future.

Internal linking and cross-referencing to external offers will be an important feature for the future electronic library. This will have to be arranged in EMIS and the journals production will have to be modified accordingly. This is combined with the need of an electronic document identifier, like it has been developed with the DOI by the commercial publishers. These activities will keep the EMIS-staff and EMIS-volunteers quite busy in the next years.

Model for an Electronic Access to the Algerian Scientific Literature: Short Description

Bakelli Yahia

Researcher on Scientific & Technical Information
CERIST research Center
03 Rue des freres Aissiou, Ben Aknoun, Algiers (Algeria)
Tel: 213 2 91 62 04-08
Fax: 213 2 91 21 26
bakeli@tassili.cerist.dz

Within the Digital-IST project (engaged by the CERIST since 1998) we try to study a mechanism of the introduction of the PC, Multimedia and Internet in the Algerian scientific publishing system. And how the Electronic Publishing (EP) technology can be exploited in the way of better access and organization of the national scientific literature. One survey among 130 higher education teachers and researchers and a systematic analysis of academic websites [1] were done. As output we establish that:

a) The national academic computer park is more and more growing because of the opened computer trade, since the latest of the eighties. Also the Internet connection rates of national scholars, research centers and Higher education institutions are growing every days. More particularly when the National Provider (CERIST research Center) ameliorated its equipment's and sales. Under these new facts a "desktop publishing" technology is more and more introduced by scholars as a mean of their knowledge producing. Also journals publishers are less and less accepting manuscripts handly written.

b) Electronic literature forms were already produced and are available through many channels such floppy disks, CD-ROMs and Internet sites. These forms consist essentially on electronic Journals, Conferences, databases, laboratories publications and other informal communications.

c) Though of these facts, national producers of scientific publications use these technologies without EP prospect. For example texts are written by softwares (such MSWord, WordPerfect..) but in order to print them with better quality and not in order to facilitate its manipulation by the publisher or the reader. And floppy disks are used just for the storage of the text and not necessarily as a distribution channel to the publisher or to be electronically read by the reader.

d) the most important electronic literature is produced by short number of producers: the CERIST (Research Centre on STI), CDTA (Research Centre on Advanced Technologies), situated in north Algeria, University of Sidi Bel Abbes (West of Algeria) and University of Annaba (East of Algeria).

e) That e-journals available on the Web are for their majority an abstracted version of printed journals. Only one is with full text form.

f) that even with the use of the technology tools, national scientific producers still practising the same writing anomalous practised in the print context. It consists on the lack of standards and legal mentions in their papers (abstracts, keywords, identifiers,...).

However within the "Digital-IST" Project we start with the assumption that even if the EP prospect is absent in producers mind, as long as the text is electronically written this means that it can be easily integrated to one EP process. We try to argue this by three kind of applications:

a) The first one consists on automatic generation of the text Head in respect of Dublin Core standard. Something like a CIP (cataloguing in Publication) in the printed books, which is never observed by national publishers. This can be illustrated by the experience done with the "RIST: revue d'information scientifique et technique", journal published by the CERIST research centre (www.cerist.dz/cerist/RIST/RIST.HTM). see fig. n°1.

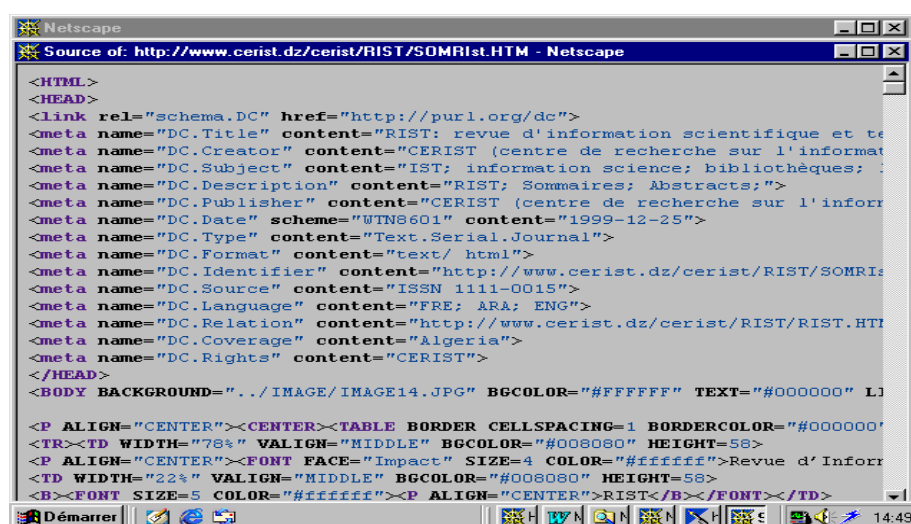


Fig. 1.

b) The second one consists on a building of a fulltext database of scientific literature. This system (which is running under WinISIS environment) gives possibility to enter a request (fig.2a), select the bibliographic record (fig.2b) and call the document (fig.2c). This is very important in our context.

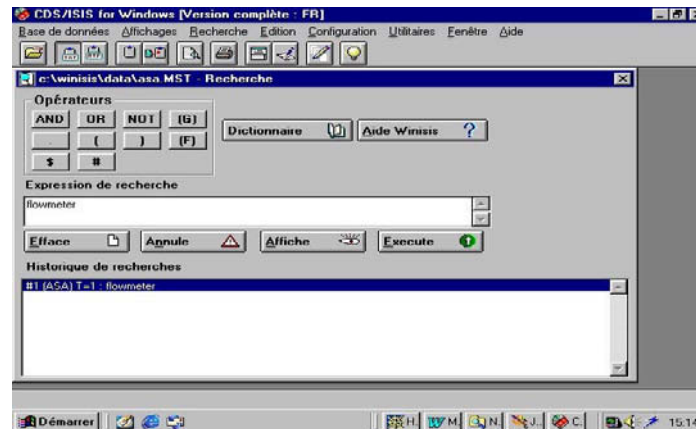


Fig. 2a.

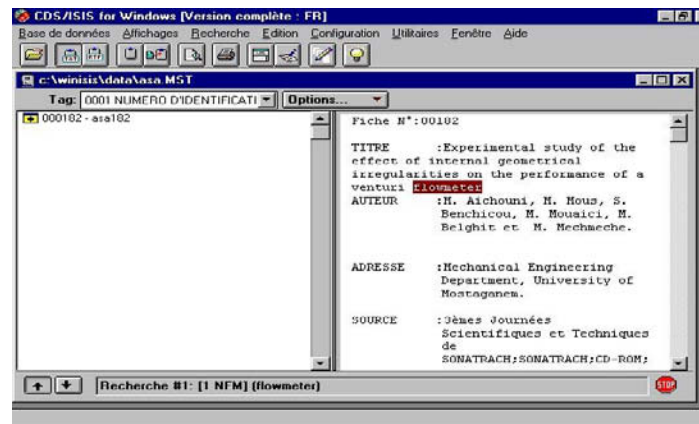


Fig. 2b.

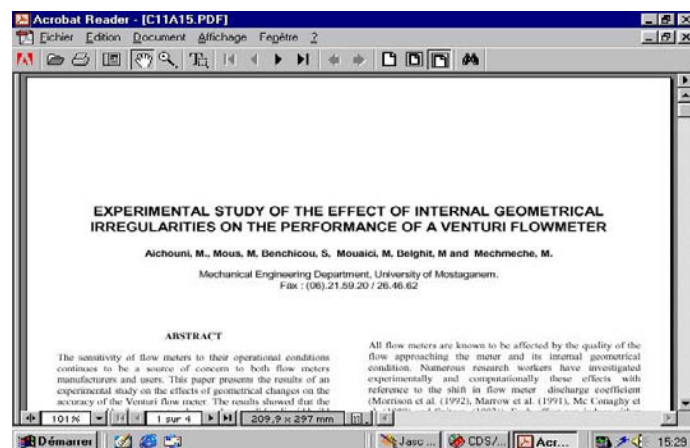


Fig. 2c.

c) The third solution consists on one Hypertext based system which is presented as a portal website to all electronic publications located on academic websites (fig. 3). So with Internet connection we have possibility to brows through national scientific websites contents even if we haven't their URL. This browsing can be done both by kind of content (e-journals, e-conferences, e-grey literature, databases...) and Kind of producers (Higher education institutions, research centres, academic publishers, learned societies...). If we take the example of the journal article, this system gives possibility to open the e-journals section as a category of publication. As a second step the user can select, among bouquet of e-journals, the title, which contains the article. Then he selects and calls the fulltext of the article from the Table Of Content.

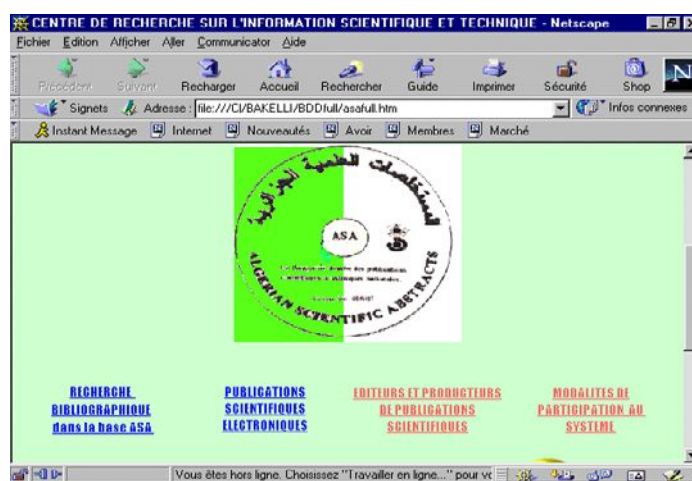


Fig. 3.

Also possibility is given to every author or publisher to put his publication in this hypertext system simply by using the electronic mail. (Fig. 4)

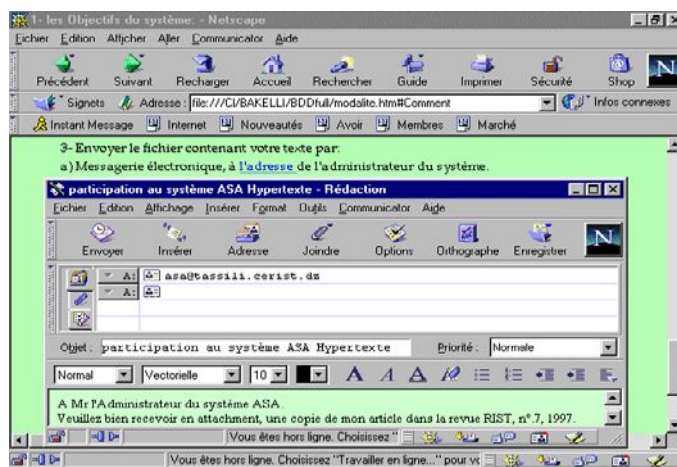


Fig. 4.

Even if this system is concerning the management of the electronic literature at the national level we argue that it can be adopted by every scientific institution (or publisher) in order to organize its own and local publications circuit and channel.

Several outputs are targeted through a wide use (among local academic institutions) of this model:

- Better access of local users to local publications;
- to diversify Preservation means of local production;
- Better visibility of local literature;
- Give publishers an infrastructure to test their electronic products;

In addition to these technical tools authors and publishers must be informed that there are other kind of tools, mainly methodological one's which facilitate promotion of EP. In this way we must mention those of ICSU [2] and INASP [3]. Also we have to attract attention to "the manual on using the World Wide Web for scientific publishing: a tool for publishers in developing countries" devoted to all third world publisher who plans to undertake publishing of electronic material [4].

References

- [1] which are available through the two main homepage of Algeria (at: www.cerist.dz) and the Higher Education and Scientific Research Ministry (www.mesrs.edu.dz)
- [2] ICSU Guidelines for scientific publishing. -Third Edition, ICSU Press, 1999.- 30p.
- [3] See <http://www.inasp.org.uk>
- [4] A manual on using the World Wide Web for scientific publishing: a tool for publishers in developing countries.-Available at : http://citd.scar.utoronto.ca/Epub_manual/index.html.

A Digital Library of Native American Images

Elaine Peterson

Montana State University
MSU Libraries, Bozeman, MT 59717 U.S.A.
elaine@montana.edu

Abstract. This paper summarizes the organizational and technical issues involved in creating a digital library of Native American images. Initial participants include a museum, an archives, and three university libraries. Using Oracle software, the shared images now constitute a database searchable by subject, date, photographer/artist, tribe, geographic location, and format of the material. Dynamic links are provided to the textual collections which house the physical images. The database resides at: <http://libmuse.msu.montana.edu:4000/nad/nad.home>.

1 Introduction

From 1998-2000 Montana State University (MSU) received from the United States federal agency IMLS (Institute of Museum and Library Services) a \$138,000 National Leadership grant to build an image database of Native American peoples that would be searchable on the Web. Included in the funding were monies for user education through the annual meeting of tribal college librarians. An initial partnership was developed between the Oracle Corporation and institutional participants—three campuses of MSU, the Museum of the Rockies (MOR), and Little Big Horn College (a tribal institution).

The IMLS project is a model program of cooperation. Smaller universities and colleges usually do not have the resources available that a museum does, such as a photo curator with expertise in preservation techniques and the archival organization of images. On the other hand, museums often lack the library and information technology expertise that are usually available on a university campus. Through cooperation the strengths of personnel at each partner institution were combined to create the database of over 1,500 images of Native Americans. The project has broadened access to new constituencies because students, researchers, and the general public now have free, direct access to important primary source material on the Plains Indian cultures previously available only by travel to Montana and the western United States.

2 Digital Library Requirements, Models, Systems, and Frameworks

The project addresses many of problems faced today by libraries aspiring to become digital. One of the most crucial areas is the software needed to run a smoothly functional relational database with full-indexing and flexible retrieval options. To accomplish this, a partnership was developed between the Oracle Corporation and the institutional participants.

On-site scanning physically took place at each location using an Agfa Scanner so that the materials never left their home site. This was important since many institutions are unable or unwilling to ship valuable materials to another location. Or, they cannot afford a high quality scanner for in-house work. Two scans of each image were made, once at a resolution of 150 dpi, described as photocopy quality (high enough for research, but not high enough for publication), and then saved in compressed JPEG and GIF files for web access from the server. The image was scanned again at 600 dpi and saved as TIFF files to be burned onto a CD-ROM as a preservation copy, which is stored off-line.

Objects were then cataloged and indexed for the database by Indian tribe, reservation, location, artist/photographer, date, format, and by 5-10 appropriate subject headings for each image using the Dublin Core. Users of the database can type in directly a subject search, or can use a simple pull-down menu for a listing from which to choose. Indexes are tied to the individual scanned images. Dynamic links exist from each image to the larger paper collection in which the image physically resides.

3 Electronic Publishing and Economic Issues

One of the reasons this project was considered for an IMLS grant was the economic issues that are addressed by the undertaking. Resources were pooled from the participating institutions. All institutions contributed images to the database. The technical expertise and equipment is housed at Montana State University. The Museum of the Rockies contributed archival assistance from their curator and photo archivist. Each of the technical and archival issues could not have been resolved by any one of the institutions by itself. In particular, smaller institutions with valuable images to contribute to the database normally do not have the archival or technical expertise on staff, nor could they afford the necessary equipment. Because of cooperation between the institutions, the bulk of the grant money could be spent for high-end hardware and software.

Additionally, similar or duplicate photographs were easily identified since resources were combined. When work occurs between institutions which focus on a single, topical subject area, overlap of collections is often inevitable. Collecting images into one database assisted all by creating one meaningful collection as a new entity.

Although the images are pooled into a central server and managed by Montana State University, each institution maintains copyright/intellectual property rights to the images that it has contributed. Instructions about copying and copyright are included on the Web pages for users of the Native American database. Because of the resolution of the images displayed on the Web, electronic watermarks are currently not being used. However, a small copyright statement is placed on the bottom of some of the more popular images.

4 Networking and Distributing Issues

The initial five participating institutions pooled their resources to create one database on a Sun Enterprise server at the MSU campus. Each institution was responsible for selection of its images, using very broad guidelines. The original primary source materials include photographs, stereographs, glass plate negatives, ledger drawings, sketches, and handwritten treaties that represent the cultures of fourteen tribes of the Great Plains region.

Currently there are over 1,500 images in the database. In addition to Little Big Horn College there are six other tribal college libraries/archives on reservations in Montana, all of which will be invited to borrow the scanner, scan their images, and FTP the files to the central server. The database has excellent potential for growth. The project may also serve as a model to agencies that are looking to undertake a similar project with another topical focus. Through the use of Oracle software, the data can be easily transferred to ASCII format or other programs for data management such as Microsoft Excel.

5 Social Implications and Issues

An important social issue faced by the project was cultural sensitivity to and use of Native American images. Images were first selected using the broad criteria of quality of the image, subject matter, and date (pre-1940). Although quality of the image was a criteria, a few images of unique items were included despite their resolution. A conscious decision was made not to enhance the images with the scanner, therefore preserving a digital copy as close to the original as possible.

General criteria were developed to encompass those images that might be culturally sensitive. All images were reviewed by a Native American consultant who marked items that might be culturally sensitive or inappropriate for widespread viewing on the Web. Material reviewed was sometimes not appropriate because it was inauthentic, or should not be shown for religious/cultural reasons. Finally, copies of the images were sent to tribal historians for further review before posting on the Internet. High quality images are available but not for commercial purposes (such as t-shirts, prints or postcards). The scanned images were made with different resolutions to allow for research on the Internet, but not for high quality reproduction.

6 Technical Information

Address: <http://libmuse.msu.montana.edu:4000/nad/nad.home>

Type of indexing: Dublin Core

List of software:

- Oracle 8.0.5 Database software
- Oracle Application Service 4.0.8.1 Enterprise Edition
- Oracle Developer 6
- Oracle Developer Server
- Adobe Photo Shop 5.0
- Equilibrium Debabilizer Pro 3.0
- OmniPage Pro 8.0
- Adobe Framemaker +SGML 5.5
- Adaptec Toast 3.56

List of equipment:

- Sun Enterprise 250 Server running Solaris 2.7
- UPS and Tape Backup Drive
- Macintosh G3 with 19" Viewsonic Monitor
- Dell Pentium II with 19" Viewsonic Monitor
- Agfa T2000XL scanner with FotoLook 3.03
- Iomega 2gig Jazz Drive
- Yamaha Rewritable CDROM Drive

7 Conclusion

Collaboration between libraries, museums, and archives will be critical to create meaningful and complete digital libraries. While some large universities are creating databases of images, this project demonstrates that there is a workable model which can be used at smaller campuses and archives.

China Digital Library Initiative and Development

Michael Bailou Huang¹ and Guohui Li²

¹ Health Sciences Library, State University of New York at Stony Brook
Stony Brook, New York 11794-8034, USA
mbhuang@notes.cc.sunysb.edu

² The Library, Capital Normal University
Beijing, 100037, China
ghlee@263.net

1 Introduction

The construction of digital library systems and digital library resource databases by utilizing advanced information technologies has become not only a tremendous challenge but also an important opportunity to information enterprises and large-scale information resources collection centers such as national libraries in various countries. Over 20 countries and regions worldwide are actively involved in building digital libraries. China is no exception. Supported by the State Council and Ministry of Culture, China's digital library project is underway in earnest. This paper aims to present an overview of current research and development of the digital library in China. It will discuss in particular the preparation work and construction of the digital library resource databases and analyze critical problems needed to be solved in the construction of China's digital libraries.

2 Overview of Current Research and Development on the Digital Library in China

The National Library of China has been following the tracks of research and development of the digital library worldwide ever since 1995. While visiting the National Library of China on October 2, 1998, Vice-Premier Li Lanqing pointed out that the developmental model of future library would be digital library and phase two of the National Library expansion project should be built into a digital library. He further suggested that the construction of digital library should adopt new thinking. [1] According to his remarks, the National Library of China put together a huge amount of manpower and funds, and dealt with numerous impediments in network, computer, software, and data processing with the cooperation of ministries related and various departments in Beijing Municipal City. The Chinese government placed digital library projects among other national level research projects. Various research institutes, libraries, universities and individuals have been conducting research on the digital library. So far, there are over a hundred research papers being published in China on topics ranging from definitions of digital library, models, systems and framework, digital library requirements, to social implications and legal issues. Some notable digital library research plans and experimental projects include: the Chinese

Pilot Digital Library proposed by the National Library of China, the Digital Library Research Plan in Tsinghua University, the Digital Library Construction Plan of Shanghai Jiaotong University, the Experimental Digital Library Project of Shanghai Library, and Liaoning Province Digital Library Project. [2]

3 Preparation and Construction of the Digital Library Resource Databases

The construction of resource databases is a major component in digital library projects, which will cost a vast amount of manpower, and material and financial resources. It's a gigantic and systematic project that extends across various industries and departments. In organizing such a construction, we should give government coordinating function into full play so that unified planning unified standard, combined construction, and resource sharing can be achieved. The authors think the construction should start with the following:

3.1 Macro-Planning

The construction of digital library resource databases first needs macro planning, that is to say, to plan the major contents of the resource database. China is a country with an ancient civilization of five thousand years. It has far-reaching social significance to sift our rich cultural heritage and process them to digitized resource database for disseminating Chinese civilization and carrying out quality education of the citizen.

3.2 Implementation

While organizing content construction of resource database, we can first select a topic, then post the content and requirement to the web in order to attract the owner of related resources to develop database jointly. The developer should develop resources with its own distinctive features. Digital library resource has international characteristics, so China's resource construction should emphasize Chinese characteristics and avoid repetition of other countries in science, technology and education fields.

4 Problems Needed to be Solved

In this section, we will analyze the following critical problems that need to be solved in the construction of the Chinese digital library project:

4.1 Understanding of the Characteristics of Digital Library Project

Chinese digital library project is a project that organizes resources in distributed computer systems, various network environment and compatible software system in proper order by the national information collection and storage center. The information in the digital library will serve regional and trans-regional clientele. The basis of digital library is user's utilization of orderly resources. Therefore, the Chinese digital library project is a project of unified planning, unified standard, and combined construction.

4.2 Government Funding

Since digital library construction needs a considerable sum of money, ideally, a large portion of the construction fund should come from the government, while a small part comes from private donations. However, digital library investment in China is insufficient compared with that of other countries.

4.3 Utilization of Mature Technology

It's better to use mature technology and achievements in scientific research in various fields in the world to speed up the process. Nevertheless, the utilization of major software should be based on the research results of Chinese computer scientists.

4.4 Markup of Object Data

After literature is digitized, how to markup object data and how to industrialize the markup are the keys to processing speed in digital library resource construction.

4.5 Cooperation of Information Collection and Storage Centers

The main function of digital library is the proper organization of resources. Such organization may happen in a number of information collection and storage centers. For each subject, the information collection and storage center needs to consider: which resources should be used, how to compile these resources, what is the aim of the project, how to reflect the theme, how to do, and so on. Only this kind of organization and coordination can avoid repetition of resource database construction and assure quality.

4.6 Virtuous Circle of the Fund

One of the keys to build a digital library system in any locale is how to acquire and operate fund so that the fund will have a virtuous circle. Many information collection and storage centers are unfamiliar with the operation of the market. Even with the initial fund, but how to make it circle is even more unfamiliar to them. Therefore, we

should seek partners who are experienced, financially sufficient, who have stable clientele to build digital libraries jointly so that the digital library project can enter into market mechanism ever since the beginning stage of the project.

4.7 Standardization

Since the construction of resource databases depends on the cooperation of various units, standardization has become an important issue. It will directly affect the quality of resource databases and the effectiveness of searching. Standards such as metadata standard, object data construction standard, data navigation standard, and so on need to be formulated as soon as possible. [3]

4.8 Legal Issues

Like their counterparts around the world, Chinese researchers and developers of the digital library also face the issues of intellectual property and copyright. Digital libraries raise more difficult and complex copyright issues than traditional libraries. However, there are no clear and perfect solutions under the framework of current law.

5 Conclusion

The development of digital libraries in China is still at the embryonic stage. With continuing enhancement of people's awareness, active participation of all types of libraries, and improvement of China's information infrastructure, digital library will become a reality in China in the near future.

References

1. Liu, X., Shen, W., Zheng, X.: An Analysis of Chinese Digital Library Project. China Computer World. (1999) 2-19.
2. Wang, B., & Meng, G.: Digital library: Its Definitions and Impacts on Traditional Libraries. Proceedings of the International Conference on New Missions of Academic Libraries in the 21st Century, Beijing, China, (1998) 462-472.
3. Mo., S.: On Theory and Practice in Building China's Digital Libraries. In: China Society for Library Science (ed.): In the Turn of the Centuries: Retrospect and Prospect of Libraries, Peking Library Press, Beijing, 1-9.

Appropriation of Legal Information: Evaluation of Data Bases for Researchers

Céline Hembise

ERSICO, Université Jean Moulin - Lyon 3
14, avenue Berthelot - 69007 Lyon - France.
hembise@univ-lyon3.fr

Abstract. This paper shows results of a study led as part of a project of the Region Rhône-Alpes entitled “Textual engineering and digital libraries”. The aim of this survey is to update appropriation behaviour of legal text by researchers in law in order to create a workstation own to the jurist, and permitting him to access and to appropriate easily a digital text. Online services of legal information are still little used by researchers and have to be developed to help users for the interrogation of online information.

1 Introduction

Legal sciences constitute a particular field of research very formalized. The legal world is marked of corporatism, always present, which obliges the compliance of social conventions and rules of functioning. The traditionalism more sticks habits and customs within the community of the jurists. Moreover, the legal text covers itself characteristics own to its speciality : specific vocabulary, anaphora, consolidated text (referring to multiple references), particular mode of reasoning,... As field of research founded on text mainly written, law is therefore a discipline very interesting to study, particularly since the access to legal information undergoes transformations due to new technologies of information and communication. So that, the jurist has to adapt himself to new tools put at his disposal. After the Videotex terminal then the CD-Rom, it is now with the Internet to gather the legal data bases. In France, legal information is initially accessible on paper then on CD-Rom and Videotex terminal. The legal data bases on the Internet then start to become extensive. The French authorities estimated that it was their duty to ensure the founding but also the diffusion of the legal data bases. Two providers share the legal information market: the public service Jurifrance and Lamy editions (with private status) that place at disposal the consultation of their own collection. Some interrogations emerge from the use of these new tools reappraising the access to information, its production and its circulation : which guidances take to place at the disposal of updated and efficient legal information, satisfying the user ? Which digital tools adopt to facilitate the legal document retrieval from digital data bases ? So, the aim is to lead to more pertinent information and definition of the existing legal services in order to improve the activity of research of the jurist. We tried to evaluate three services suggested to the jurists: Lamyline, the Jurifrance station and the server Internet Jurifrance.

Thus, the goal of this experimentation is to estimate the use of the legal text available on digital data bases in order to found an efficient workstation peculiar to the jurist.

2 Our Experimentation

This work follows upon a pre-survey conducted near some jurists in connection with their use of the legal data bases. The study had to claim to define the various access modes to the legal documents by the researchers in law and to establish their method of appropriation of those texts. The goal was to try to fix their documentary and intellectual practices in order to determine their cognitive process at the time of a textual appropriation. This survey shows clearly that the use of those legal servers by researchers in law is yet only in its early stages and that the information flow still remains problematic. However, some tendencies start to take shape as for their practices. From primarily individual works, the researcher has a propensity to remain shut away in his usual world. Not very inclined to the exchanges with people considered as “outsiders” from his community, he preserves a corporatist spirit experienced by his process of information, confining himself to the question of his/her colleagues. In this situation, the restraints raised with the use of new technologies indicate the very strong importance of paper, book above all other digital support. The jurist relatively seems recalcitrant to the use of the Internet that he considers more as a waste of time for the information retrieval than an asset. Another fear re-appears regularly: non reliability of the legal data. The researchers reappraise the data put on line trusting the diffusers asserting of no technical and legal guarantee as for the validity of the digital data. This fear of the tool is based on the fact that the researchers don’t know the Internet, cannot make use of it. A priori, some would be ready to use it provided that a training is given to them.

A second investigation was carried out near the doctors in law and researchers in their experiments of two digital legal services : Lamyline and Jurifrance. The matter is to evaluate their relevance and their efficiency by the practice of the participants in direct mode. The project is to collect data on ergonomics of the product tested, its functionalities, its user-friendliness and if it meets or not the requirements and needs of the potential user. The results got aim to establish some recommendations close to designers of those products with a perpetual retroactive, that is, persistent measure of new effects induced by those improvements via questionnaires or regular survey near user, in order to have a constant staying up of their opinion and needs on the product which are dedicated to them. The underlying objectives to the postulate consisted in checking: the appropriateness of legal text on line with the request of user, the presentation of legal text (layout, facility of access), the relevance of tools and functionalities of retrieval information for the user, the adaptation of operation and of processing of legal text with needs of user, the needs and expectations of user. Doctors in law, all categories taken together, and the teacher-researchers in law of Lyon university, formed the target of this investigation. Participants had to evaluate a choice of two interfaces out of three. During the appraisal of the interfaces, the user let foresee his perception of the observed services.

However, it must emphasize that the participants don't always have an objective view of these services at the time, either they did not know the existence of these services, or they compared them with those they had already practised. The users put up high financial costs to the interrogation on line whereas for the same services, the exemption from payment is obligatory abroad (Quebec and the United States in particular). He also regrets the lack of immediate availability of the legal texts from their publication. The circulation period of the information on the Internet seems them too much long. However, the user takes as comparative element the distribution times of legal information appearing on paper. It would be necessary waiting some time so that the producer and the distributor of digital information be designed in their system and propose some services as well developed as those allowed by the paper medium. Lastly, it should be taken account the fact that the potential users are not all familiar with the data-processing tools and that a time of training and formation is necessary. This component is often forgiven by the designers - such for example the absence of an online help - and thus influences the more or less good success in the experiments of those services.

3 Recommendations

Therefore the objective is striving for the conception of a workstation suitable to the researchers because the specificity of their requests forces to think a completely adequate tool to their retrieval. In particular, the few emerging points concern:

3.1 Content of the Data Bases

- Some accessible document so much in plain text as with their summaries.
- Some legal text respect the formal presentation of the law.
- The systematic consultation of the consolidate text.
- A contents of all the supports existing.

3.2 Functionalities

- A functional ergonomics equipped with legible screens in the display (mosaic).
- An on line help efficient.
- An update of the lexicons as well as the revision of their coherence.

3.3 Recovery and Exploitation of the Results

- The possibility of modify and format an exported document.
- The direct insertion of a document in own one's file of word processing.
- The export of a list of references with the plain text.
- The instantaneous printing without including the function of recovery.

3.4 Workstation

- The creation of personal files, recovering them remotely, on the desk.
- The constitution of a virtual private library.
- The possibility of creating oneself an historic of retrievals allowing not to be lost.
- Navigating between different data bases without systematically opening them.
- The access to all the bases and legal data banks existing.

4 Conclusion

According to the user, the principal obstacles in the consultation of digital data bases are currently the excessive cost of the view of the references as well as the access modes to those documents. The development of access points to such data bases would seem relevant via public organizations, the one being able to assume such a financial cost. The practices of information retrieval of the lawyers thus remain very conventional in so far as information on the paper medium stays the prerogative of the information retrievals. But an evolution takes shape with rising generation, already practising the data-processing tool and knowing better the new technologies. However, the restraints emitted with their experiments strongly remain. The creation of a workstation - in the sense where "workstation" would indicate an interface dedicated to the user (thought for him) and offering various tools : consultation, experiment, appropriation and recovery of a document own to the jurist and would allow him to appropriate fully a document, from its access to its classification or virtual storage. But it seems that the users don't become aware of the stake which represented such tools for their work in the future. The user actually sees only his first interest, that is those services cost expensive and are rather slow, without understanding that the works on paper are expensive for a library. In term, with the development of the digital libraries, the put on line of legal services is essential but some lawyers did not become aware yet of it. They perceive neither the advantages nor the drawbacks of those products and grant little importance to it, since currently, they don't allow them to improve their way of working and to be more efficient in their activities. It seems that they did not make the assessment of the social and professional change that could generate the expansion of those new tools of work.

References

- ADIJ : L'information juridique : contenu, accessibilité et circulation, Congrès international de l'ADIJ, Paris (22-23 octobre 1998)
- Gibert H.: Lecture assistée par ordinateur et droit, Congrès ADQIJ, Montréal (1992)
<http://liguria.crdp.umontreal.ca/crdp/fr/equipes/technologie/conferences/aqdi/Congres92/G03/06/98html>
- Hembise C.: L'accès à l'information juridique par les étudiants en droit : le point de vue des utilisateurs, Colloque : Les e-modes de diffusion et d'appropriation des connaissances, 28-29 juin 2000, Lyon. <http://www.univ-lyon2.fr/SeminaireARASSH>
- Hembise C., Metzger J.P., Rey J., Rapport d'étape sur le projet ARASSH, Bibliothèque électronique et ingénierie textuelle, Lyon : Arassh (dec.1998)

Publishing 30 Years of the Legislation of Brazil's São Paulo State in CD-ROM and Internet

Paulo Leme¹, Dilson da Costa¹, Ricardo Baccarelli¹, Maurício Barbosa¹,
Andréa Bolanho¹, Ana Reis¹, Rose Bicudo¹, Eduardo Barbosa¹, Márcio Nunes²,
Innocência Pereira Filho², Guilherme Plonski¹, and Sérgio Kobayashi²

¹ Projeto IMESP/CECAE-Universidade de São Paulo
Prédio da Antiga Reitoria 7º andar – Cidade Universitária
05508-900 São Paulo, SP, Brasil
paulol@usa.net

² Imprensa Oficial do Estado – IMESP
Rua da Mooca 1921
03103 São Paulo, SP, Brasil
cti@imesp.com.br

Abstract. About 60,000 legal acts covering the last 30 years of the São Paulo State legislation are published in CD-ROM and the Internet. The effective search engine implemented yields to a swift and intuitive system using hypertext feature for easy navigation through related legal acts. It is already reflecting a significant improvement in administrative and legal procedures in government offices, legislative and judiciary departments, municipal administrations as well as private lawyer offices, corporations, universities, and members of the community.

The Legislation of Brazil's São Paulo State constitutes a major working tool for several sectors of the Brazilian society, from the Governor, who relies his administrative decisions upon it, the Legislative Assembly who elaborate new laws, up to members of borough communities who organize themselves to demand better government services.

The main source for the research on the legislation has been so far the *Diário Oficial*, the daily official government publication kept in the libraries of most institutions. This way of information searching is quite complex, time consuming, tedious, and incomplete. Modern digital information technology offers rather fast, efficient and low cost management resources. Aware of such picture, the Official Press – IMESP – publisher of the *Diário Oficial*, decided to fund a project for the publication of São Paulo State legislation in CD-ROM and in its Internet site (www.imesp.com.br), carried out by a multidisciplinary group of the University of São Paulo.

The target was about 60,000 legal acts, distributed among laws, complementary laws and executive decrees, covering the last 30 years of the São Paulo State legislation.

Except for the last 4 years in which the legal acts were already generated in digital form, the data source was the printed *Diário Oficial*, incidentally the biggest daily

publication in the whole world: from 1200 to 1500 pages per day. To obtain digital data comparative testing was done between scanning followed by optical character recognition (OCR) versus regular typing and proofreading. The choice remained in this last option by time and cost optimization criteria.

Careful evaluation of the software available in the market was done, including BRS Search, Folio, WordCrunch, DB Text, LightBase and Digital Altavista. The Folio family of products, comprised by Folio Views, Folio Builder, Folio Site Director and Folio Integrator, was the choice which gathered the best features, namely:

- easy and fast data search and retrieval, provided by the previous comprehensive indexation of the words in the database (a text base not a relational one)
- easy and fast data presentation in html (once the search is done, the html page to present the results must be assembled on the fly)
- free format database
- high compression rate
- free text and boolean search
- hypertext linking
- local support
- adequate price

Extensive typing was done using MS Word with specific style format appropriate to be exported to Folio Views. Auxiliary software was developed for data entry, filtering and classification.

To facilitate information retrieval, fields were attributed to each act, namely: its type (whether law, complementary law or decree), number, approval date, subject and author (of law project); the journal's volume, number and issue date; and the name of the governor in charge. A list of main subjects was developed in accordance with the criteria used by the library of the Legislature, as a first approach.

The final three information databases, or infobases, comprising each type of act (law, complementary law and decree) are so compact that add up to no more than one hundred megabytes.

Information retrieval may be performed by any field of a particular act, any field of the *Diário Oficial* in which it was published, or by any boolean combination of fields and words within the database. The system is easy to use, due to the intuitive structure implemented, and fast as well due to the swift Folio inherent search engine. The hypertext feature provides easy navigation through related legal acts.

In order to assist the complex task of information recovery the infobase is presently undergoing a meticulous and comprehensive librarian indexing using appropriate descriptors. This assignment will lead to a thesaurus developed specifically for this infobase thus avoiding synonymy and inadequate terminology in the classification of the legal acts yielding a rather consistent and refined information retrieval.

Immediate results of the present work is already being noticed and is expected to reflect a significant improvement in administrative and legal procedures in government offices, legislative, judiciary, civil and military departments and secretaries, as well as about 650 municipal administrations of the State of São Paulo. Lawyer associations and private offices, unions, private corporations, universities, colleges and members of the community in general are also supposed to benefit.

Two important by-products of this project are the encouragement of citizenship exercise making the legislation easily available to the population, the democratization of the information so to speak, and the transparency of the government issues.

Electronic Dissemination of Statistical Information at Local Level: A Cooperative Project between a University Library and Other Public Institutions

Eugenio Pelizzari

Biblioteca Centrale Interfacoltà, Università degli Studi di Brescia, Italy. ¹
pelizzar@eco.unibs.it

Background

The European process of integration is designing a new institutional scene involving a progressive *strengthening of international and local governments*, with a gradual weakening of intermediate levels, especially the National States. Local communities, from an economic and social point of view, will probably have a more direct relationship with each other. We predict that *inter-European competition will take place at the level of local economic and social systems*.

In recent years, the need for statistical information has increased progressively, due to the international integration of markets, public administration and ordinary people.

In an institutional landscape such as this, public statistical information will assume a double and fundamental role:

- To support public decision making during the programming and control phases;
- To offer information to the community, which anyone can draw on for their own personal use.

The *evolution of computer science technology and telecommunication systems* (Internet in particular), give new perspectives to these questions, deeply *modifying the production and dissemination processes of statistical information*.

Brescia is a town in the northeast of Italy with a long tradition in statistical data resources. At present, the local statistical collections are not easily accessible because they are scattered throughout different institutions or are located within specialized structures, not yet catalogued or catalogued with inhomogeneous cataloguing rules. Also, the available services offered are not coordinated and are partially unsatisfactory.

The institutional background and the opportunities given by the new technologies calls for the establishment of a *cooperative project* on a local level.

In Brescia, statistical collections are mainly owned and managed by four Public Institutions: The Camera di Commercio, Comune, Provincia ed Università degli

¹ I wish to thank Marco Trentini, Unità di Staff Statistica, Comune di Brescia, Italy, for his helpful comments on this paper.

Studi. Each of them, due to historical and institutional reasons, is the depository only of a part of the local statistical documentation and datasets.

It is possible to emphasize some aspects:

1. The statistical collections are complete at the territorial level (from local to international statistics) and cover a long period of time (from the present going back to post-World War II);
2. The collections have grown in an uncoordinated way; in spite of this (as a consequence of the specific mission of the different institutions) there is not much duplication of materials;
3. The organisations involved have developed different ways of organising, managing and offering the statistical information; in this sense, cooperation enables synergies and permits coordinated answers to the different demands.

The Project: Objectives, Method, and Activities

The idea of the project stemmed from the collaboration of two public institutions:

- *Biblioteca Centrale Interfacoltà* dell'Università degli Studi di Brescia (Central Library of the University of Studies of Brescia)
- *Ufficio di diffusione dell'informazione statistica* (Office for the Spreading of Statistical Information of the Municipality of Brescia)

The institutions contacts were made with and with whom it is necessary to define a collaboration premise are the following:

- * *Camera di Commercio di Brescia - Ufficio Statistica*
- * *Provincia di Brescia – Ufficio Statistica*
- * *Regione Lombardia – Ufficio Statistica della Presidenza della Giunta*
- * *ISTAT – Ufficio Regionale Lombardia* (ISTAT could be interested in supporting the Project, partially assuming the following, specific functions: supplying the statistical data and databases collected and processed by the Institute, helping define the standard elements of classifications and providing methodological support on the criteria for treating the documentation).

Other institutions could join this initial group in the future, on the basis of a common agreement.

The main objective of the project is to *supply information and support* to public administrations and to other decision-making bodies, *maintaining a large library of data files and relevant statistical documentation*. The project's main strength lies in its ability to respond to the demands from local and university communities, for statistical information through a unique WEB interface.

The above objectives will facilitate access to the statistical collections of local Institutions, regardless of where they are physically located. Related services, such as exchange of information, advising, references, and personalized processing of statistical data, will be available.

The Project is articulated on three levels:

1. Homogenous classification of all of the currently available statistical information (documentation, data files and bibliographic databases) using international standards, and making it available through Internet.
2. Implementation of services for consulting and accessing materials, such as:
 - Virtual Library of (that is a Meta-OPAC for access to) local statistical documentation;
 - WEB access to the data library;
 - Inter-Institutional Loans;
 - Electronic Document Delivery;
 - Virtual reference desk;
 - Other services.
3. allowing access to local administrative databases and datasets, after having enriched them with the relevant meta-data such as the objective of the information collected, the survey methodology, archives from which the data were drawn from, quality of the data, updating of the data, and other characteristics. Confidentiality will be respected in the statistical information provided (following the “data library” model).

Specific Objectives and Services of the Library

As we have previously specified, one of the major strengths of the project is the cooperation between different agencies to create a unique point of access to all statistical resources (bibliographical databases, datasets, documentation...) available on a local level. If it is true that only a system such as this can guarantee the success of the project, it is also true that each institution must pursue its own mission, in full respect of the autonomy of everyone. It may be interesting, therefore, to try to specify the specific objectives of the Library in this project, the specific services and relative benefits for the user. From a general point of view we can say that, through this project, the Library will be able to offer the University community access to a wide variety of numerical data resources and related services. The core function of the Library will be to support the empirical research and teaching activities of the faculties of the University of Brescia, students and staff.

With reference to its users the Library will be in a position to provide (either directly or, more often, through the contributions of the partners involved in the project), services in the following major areas:

- *Data Library*: by archiving and maintaining data files collected by University researchers the library will function as a data archive, (which will contribute to implement the general data library accessible via the Web);
- *Data Consulting*: The Library should advise users on the extraction of data and on the use of computer programs for analysis. Advice can range from helping students to extract basic data for a paper or an exam to addressing doctoral candidates or researchers to those professionals (both inside and outside the university) who can help them better manage very large databases containing

millions of records. This service will include technical assistance in creating custom subsets of just the cases and variables needed in the software dependent format required for the user statistical package. Depending on available human resources, the Library will help researchers and students with the use of statistical software in the definitions of data files as well as with the composition of bibliographic citations for computer files.

- *Data Access:* The Library will provide microcomputer workstations for WEB based access to the central System where documentation, all online data and all bibliographic and other "look-up" databases will be available. The workstations will also provide facilities for moving data across the University network and for transferring data to floppy, Compact and Zip disks for use on stand alone computers and between hardware/software platforms. Maintaining systemic contacts with the others partners the library will be able to assist users in searching for and getting copies of needed data resources available elsewhere.
- *Instructional support:* The Library will help its users to extract appropriate data for their purposes; the library staff will conduct guided introductions to the service and also in the construction of custom data sets for instructional or research use.
- *Data Acquisition:* identification and acquisition of new data files and related documentation will respond to the growing need of statistical information from the academic community for research and teaching.

Comments

At present, the Project is in the institutional review and approval phase, especially with the local administrations statistical offices and with the Ufficio Regionale Statistica of the Regione Lombardia.

Besides the countless informal meetings, a first formal meeting was held with all the prospective institutions who could be involved in the initial stage. All the participants gave their adherence in theory to the project, tied to the creation of a technical board which, in a short period, can draw up a plan, the implementation phases, costs and finances required.

The University Library will play the double role of promoting and coordinating the project. Together with the other agencies involved, the Library should be able to distribute advanced services in the statistical field, particularly those international statistics covered by any one of the other individual agencies. The benefits will be immediate for the university community (students, teachers, researchers), public administrators, enterprises and the local community. For the first time, these groups will have direct access to complete statistical information of fundamental importance in a wide range of aspects of the social and economic life.

Building Archaeological Photograph Library

Rei S. Atarashi ¹, Masakazu Imai ¹, Hideki Sunahara ¹,
Kunihiro Chihara ¹, and Tadashi Katata ²

¹Nara Institute of Science and Technology
Takayama, Ikoma, Nara, 630-0101, Japan
Tel. +81 723 72 5103 Fax. +81 743 72 5620
ray@dl.aist-nara.ac.jp, imai@is.aist-nara.ac.jp,
suna@wide.ad.jp, chihara@is.aist-nara.ac.jp,
²Tezukayama University
Tezukayama, Nara, Nara 631-8501 Japan
katata@tezukayama-u.ac.jp

Abstract. The photographs taken at excavation fields are one of most important materials. It is expected to digitize and save these photographs in the computer system because of difficulty of management to many photographs and losing color problem. We designed and implemented prototype of archaeological photograph library. The library is designed based on Dublin Core Metadata Element set and XML. We describe the photograph library project and design concept of the library.

1 Introduction

The study of ancient societies, Archaeology, in Japan examines the period from ten thousands B.C. to seventh century A.D. For archaeologists it is important to taking many photographs of these historical sites like remains, buildings and graves for the purpose of leaving precise records. The thousands of photographs have following problems.

- It is difficult to manage, classify and sort these photographs for investigation. The maintenance cost is so high, it is hard to find appropriate data or detect relation between these data.
- Old films of photographs have become to lose color. In general, films of photograph are able to keep color only twenty years. The films that have lost colors also lose the value as precise materials; it is urgently necessary to take action for saving these data.
- Most of photographs are kept in a warehousing, not used other archaeologist or another purposes.

It is one solution to problem to digitize photograph and archive to the large computer system. The advantages of building digitized film library are not only saving enormous data but also finding appropriate easily and enabling to describe cross-reference. The archaeological remains have metadata (data about data) that explain period that these are used, type of the remains and place these are excavated so on. The goal of this system is publish these data and enable to inter-connect or

exchange data to similar server, the metadata definition have to be based on international standard. The core metadata of archaeology is discussed by Archaeological Sites Working Group of the International Committee for Documentation (CIDOC)[1] of the International Council of Museums (ICOM)[2]. On the other hand, Dublin Core Metadata Element set [3] is international standard, especially, adapted by library and museum. The library is based Dublin Core because it is open to public though the Internet.

In this paper, we describe the project building film library of Japanese archaeological relics: the overview of this project, proposal archaeological metadata based on Dublin Core, design of database system and database.

2 Archaeological Photograph Library Project

2.1 Target

The emeritus Professor Katata has taken photograph at the excavation field; the collection is about tens of thousands. At the first step, the target is digitizing his entire photograph and saving in the computer organized by giving metadata. The final goal of this library is making connection to other server through the Internet to searching across these sites and exchange data. In this paper, the goal is building stand-alone server that has all facility such as scanning, database, giving metadata and browsing.

2.2 System Design

The Archaeological Photograph Library consists of three four components: digitizing, database, giving metadata and browsing.

In digitizing component, the films of photograph are scanning to digitized data such as JPEG or photoCD format. The giving metadata component consists of two tools: metadata input form and input support tool that assists people who give metadata.. To retrieve and browse is realized through WWW browser. The data retrieval is mainly executed using metadata as a key.

3 Designing Metadata of the Archaeological Photograph Library

Dublin Core Element Set consists of fifteen core elements, each element is optional and repeatable. It is possible to add sub element called qualifier. Definition of qualifier in the specific field is discussing for saving flexibility and interoperability.

Table 1 shows Metadata designed to this photograph library. Since this list is provisional version it might be change by implementation and evaluation. After the final version is completed, we will propose the designed Metadata as an archaeological standard.

DC.Title	Title of the digitized data.
DC.Title.relic	Title of the relic.
DC.Title.ruin	Ruin the relic belongs.
DC.Title.historicalsite	Historical site the relic belongs.
DC.Creator	
DC.Creator.relic	Person/people who make the relic.
DC.Creator.photo	Person who take a photograph of the relic.
DC.Creator.digitize	Person/organization who digitize photograph.
DC.Subject	
DC.Subject.relic	Keywords of the relic
DC.Subject.ruin	Keywords of the ruin
DC.Subject.historicalsite	Keywords of the historical site
DC.Subject.photo	Keywords of the photograph
DC.Description	Description of the relic.
DC.Publisher	Person/organization who publish the digitized photograph.
DC.Contributor	Contributor.
DC.Contributor.excavation	Person/organization who contributed to excavate the Historical site.
DC.Contributor.photo	Person/organization who contribute to take photograph.
DC.Date	
DC.Date.period	Era the relic was used.
DC.Date.excavation	Date the relic was excavated.
DC.Date.photo	Date the relic was took photograph.
DC.Date.digitize	Date the photograph was digitized.
DC.Type	
DC.Type.period	Period the relic was used.
DC.Type.relic	Type of relic (earthenware, sword etc.).
DC.Type.ruin	Type of ruin (house, grave etc.).
DC.Type.photoangle	Angle of the photograph.
DC.Format	Format of digitized data.
DC.Format.size	Size of digitized data.
DC.Identifier	Identifier of the digitized data.
DC.Identifier.address	Address of the Historical site.
DC.Identifier.point	Longitude and latitude.
DC.Identifier.number	Number of photograph.

DC.Identifier.relic	Number of relic in the photograph.
DC.Source	Relic (object) in the photograph.
DC.Language	Language
DC.Relation	
DC.Relation.relic	Relation among other relics.
DC.Relation.panorama	Panorama of the relic, ruin, historical site.
DC.Relation.neighborphoto	Photograph came out neighbor of the relic.
DC.Coverage	
DC.Rights	People/organization who hold rights.
DC.Rights.relic	People/organization who hold rights of relic.
DC.Rights.photo	People/organization who hold rights of photograph.
DC.rights.digitize	People/organization who hold rights of digitized data.

Table 1. Metadata for archaeological photograph

4 Database and Browsing

The database is a core component in the library system. Generally, the most important function required for database is searching appropriate data fast and exactly. The database is also expected to develop for providing support to person who gives Metadata automatically by learning. There are two ideas to implement this function. One is implementing a reference counts in the database. Metadata that used many times might be used next time. This method is not fair because the number of appearance dose not indicate the importance of information. Second is making relation table to select appropriate Metadata. It is more certain than the first method, since it can describe to original relation and apply to select showing Metadata.

The library opens to the public through the Internet with WWW browser. The format is XML to display result of searching, digitized film data and metadata. It makes available to exchange data among other server or library in the futer.

Conclusion

We described importance of digitizing and archiving of archaeological photograph. The design of the archaeological photograph library is completed and prototype implementation is started. In the future, the evaluation process is required for propriety of the metadata based on Dublin Core Metadata Element set and database.

References

- [1] International Committee for Documentation (CIDOC) (<http://www.cidoc.icom.org/>)
- [2] International Council of Museums (ICOM) (<http://www.icom.org/>)
- [3] Dublin Core Metadata Initiative (<http://purl.org/dc/>)

EULER – A DC-Based Integrated Access to Library Catalogues and Other Mathematics Information in the Web

Bernd Wegner

Technische Universität Berlin, Germany
wegner@math.tu-berlin.de

Abstract. Literature databases, scientific journals and communication between researchers on the electronic level are rapidly developing tools in mathematics having high impact on the daily work of mathematicians. They improve the availability of information on all important achievements in mathematics, speed up the publication and communication procedures and lead to enhanced facilities for the preparation and presentation of research in mathematics. The aim of this article is to give a more detailed report on one of these projects, the so-called EULER-project, developing a search engine for distributed mathematical sources in the web. Main features of the EULER deliveries are uniform access of different sources, high precision of information, de-duplication facilities, user-friendliness and an open approach enabling participation of additional resources. The partner of the projects represent different types of libraries and moreover different types of information in the web. The functionalities of the EULER-engine will be described and a report will be given on the transition from the prototype developed in the project to a consortium based service in the internet.

1 Synopsis

The aim of the EULER-project is to provide strictly user-oriented, integrated network based access to mathematical publications. The EULER-service intends to offer a "one-stop shopping site" for users interested in Mathematics. Therefore, an integration of all types of relevant resources has to be a goal of such a project: bibliographic databases, library online public access catalogues, electronic journals from academic publishers, online archives of preprints and grey literature, indexes of mathematical Internet resources. They have to be made interoperable, using common Dublin Core based Metadata descriptions for example. A common user interface - which will be called the EULER-engine - has to assist the user in searching for relevant topics in different sources in a single effort. As a principle, the EULER system should be and has been designed as an open, scaleable and extensible information system. Library users and librarians from mathematics in research, education, and industry are the main participants of such an enterprise.

EULER is an initiative of the European Mathematical Society, and especially focuses on real user needs. Standard, widely used and non-proprietary technologies

such as HTTP, SR/Z39.50, and Dublin Core are used. Common resource descriptions of document-like objects enable interoperability of heterogeneous resources. The EULER-project develops a prototype of new electronic information services. Hence most relevant information of one subject area (mathematics) is integrated in this project (one-stop-shopping). The EULER-results have been designed in such a way that they are easily portable to other subject domains.

Users are enabled to make effective use of the mathematical library-related information resources offered with a single user interface. Time-consuming tasks associated with the use of non-integrated services have been eliminated. The user has been enabled to search for and localise relevant documents. In many cases he can retrieve the full text of an article electronically.

The Project is funded by the EU within the programme Telematics for Libraries.

2 Objectives and Structure of the EULER-Project

As mentioned above the aim of the project is to provide strictly user-oriented, integrated network based access to mathematical publications, offering a "one-stop shopping site" for users interested in Mathematics. Therefore, an integration of all types of resources mentioned above is necessary. Since EULER combines descriptions of resources (bibliographical databases) with the complete text of documents, free resources with commercial ones and databases with very different structures, retrieval systems and user interfaces, this integration had to built upon common resource descriptions. This glue or intermediate level is accomplished by using descriptions of all resources following the Dublin

Core (DC) metadata standard, recently developed and published as an Internet draft.

Technically, the integration of the different resources has been accomplished by producing DC metadata for all resources (by means of conversion, automatic generation or metadata creator software), and collecting it into front-end databases for every individual EULER-service. A retrieval and search software, the EULER-engine, uses these metadata databases as sources for a distributed search service. The integration approach is based on the Z39.50 standard or on HTML-form based data interchange.

At distributed servers, multilingual EULER-service interfaces are provided as entry points to the EULER-engine, offering browsing, searching, some document delivery and user support (help texts, tutorial etc.). The interface is based on common user friendly and widely used web browsers (public domain or commercial) such as for example Netscape. The (multi-lingual) user interface has the common features of every good Internet service and a self-explaining structure. The user has one single entry point to start of his information search. The searching contains

- a subject oriented browsing

- a search for authors, titles and other relevant bibliographic information
- a subject oriented search in different information resources.

Full access to the implementations of the project results is available at all participating libraries (SUB, UNIFI, NetLab, CWI), and in a regional network of French research libraries (co-ordinated by MathDoc), tailored to specific institutional needs. Restricted demo access is available for the general public. The European Mathematical Society encourages European mathematicians from research, education and industry to use and evaluate the new services. Overall scientific quality of the services are secured by the appropriate Committees of the European Mathematical Society.

Practically the main objectives of the project correspond to a set of work-packages. An initial Requirements Analysis work-package covered user requirements, final discussion and definition; revision of methodologies, test and evaluate alternative concepts for the EULER system; the integration of new relevant developments in the EULER system; standard developments monitoring, observing the developments of new important relevant standards, participation in relevant standard definition discussions.

The Resource Adaptation work-package builds the basic set of EULER Metadata Databases that are finally accessible from the EULER-engine like scientific bibliographic databases, library OPACs, preprint servers, peer-reviewed electronic journals, mathematical Internet resources. Bibliographic databases and OPACs cover the broader scenario of automatic metadata to metadata conversion. Peer-reviewed electronic journals, preprint servers, and mathematical Internet resources cover the broader scenario of resources harvesting, metadata creation (automatically or manually), and access to networked resources.

The EULER-engine Implementation work-package - carried out in parallel to the Resource Adaptation work-package - has designed and implemented the EULER-engine. The EULER Engine acts as an "intelligent" gateway between users and the metadata databases produced in WP-2 by providing:

- user oriented interfaces and help tools,
- the capability to re-map searches and browsing to the metadata databases,
- the capability to collect answers (i.e. hits) and to present them by ranking, filtering, ordering etc.

This includes both the user interfaces and the interfaces to the partners metadata databases and other selected Internet resources.

During the Evaluation and Demonstration work-package - to be carried out after the release of the EULER Engine (beta version) in July 2000 - selected groups of users will start system evaluation. The work intends to measure the system suitability and scalability and the satisfaction level of users with the service.

The last work-package is Information Dissemination and Exploitation Preparations. Information dissemination took place and will take place via

professional journal articles, presentations at conferences, and similar events. Relevant reports of the project are made publically available on the World Wide Web. The final exploitation plan for EULER services and other project results is under discussion. Commercial exploitation for future operation of EULER services and transfer of EULER results to other subject domains is under consideration.

3 The EULER Partners

The currently accessible contents for the EULER-engine are provided by the partners of the project. The group includes libraries spread out all over Europe, who represent also different types of libraries.

The State Library of Lower Saxony and University Library of Göttingen (SUB Göttingen) represents a library with national responsibility to collect all publications in the field of pure mathematics. It is one of the five largest libraries in Germany. Göttingen is in charge of more than 20 specialist collections supported by the German Research Association. The CWI library (<http://www.cwi.nl/cwi/departments/BIBL.html>) is the typical candidate for a research library of a national research center, - CWI in this case. It has a large and extensive collection of literature in the fields of mathematics and (theoretical) computer science. The University of Florence as a project partner represents the typical university library with its distributed department libraries. The libraries automated management of the University of Florence started in 1986 with the participation to the Sistema Bibliotecario Nazionale (SBN), promoted by the Italian Ministero per i Beni Culturali ed Ambientali. Currently, 50 libraries are, spread over Florence, including those of faculties, departments and institutes.

A partner specialised in digital libraries and net-based information is represented by NetLab. The name stands for the Research and Development Department at Lund University Library, Sweden. It is running or participating in a number of projects in collaborative efforts with other institutions and organisations from the Nordic Countries, Europe and USA. DESIRE (<http://www.ub2.lu.se/desire/>), the Development of a European Service for Information on Research and Education, is one of the largest projects in the European Union Telematics For Research Sector of the Fourth Framework Program. In addition, MDC as a national center for co-ordination and resource-sharing of mathematics research libraries and mathematics departments is representing libraries and library users in EULER. MDC stands for "Cellule de Coordination Documentaire Nationale pour les Mathématiques" (MDC).

Together with the European Mathematical Society and the Heidelberg Academy of Sciences FIZ Karlsruhe provides the longest-running international abstracting and reviewing service for mathematics, Zentralblatt MATH. Zentralblatt MATH (<http://www.emis.de/ZMATH>) covers the entire spectrum of mathematics and computer science with special emphasis on areas of applications with about 70.000 items per year. Development efforts have been undertaken in co-operation with MDC to offer enhanced search functions in the MATH database via the World Wide Web.

Special links to electronic articles and library based document delivery services are offered with the database searches. The project is an initiative of the European Mathematical Society (EMS), which represents the community of library users interested in mathematics from the whole of Europe. The EMS will bring in its Electronic Library of Mathematics, distributed through EMS's system of Internet servers, EMIS, <http://www.emis.de/>. This Electronic Library is today the most comprehensive archive of freely available mathematical electronic journals and conference proceedings.

Experiences from prior and on-going work of these institutions, sketched above, form the baseline of the EULER-project. They cover more or less all aspects of knowledge which will guarantee that the project will lead to an excellent product. Some of them already agreed to be part of the consortium which should take care of the permanent EULER-service, once the EULER project will have been terminated.

4 Current Achievements of the EULER-Project

The evaluation of the user needs had been finalised in the first period of the project. On this basis the first draft of the EULER-engine, the adaptation of resources and the user interface had been developed. In the middle of 1999 the whole system had been offered for the alpha-test to a broader community of experts and users in order to get comments for improvements and extensions. The response from the community was positive in general. A lot of specific proposals have been obtained to improve the usability of the system. The EULER-engine and the metadata-maker are working in a reliable way. The unique identification of sources and the de-duplication check lead to quite homogeneous search results offering comprehensive information within the hit list.

These recommendations and additional results from internal discussions between the EULER-partners were taken into account for the development of the beta-version of the system. In particular special efforts will be spend for improving and extending the user interface. Special selection facilities between resources and more options for prescribing general ranges for the search will enable the user to get quicker and more precise results with the EULER-engine. The test of the beta-version will be carried out in the middle of 2000.

The adaptation of resources has been finalised in parallel to the work on the beta-version. As an additional test a restricted set of resources has been invited to adapt their content to the requirements for being searchable within the EULER-system. They have offered some support from the EULER-group to get the adaptation work done. This lead to an improvement of the preprint information available within the EULER-system. Some other modest extensions of the accessible content are in preparation.

The detailed functionalities provided by the EULER-system through its current user interface can be checked directly under the URL

<http://www.emis.de/projects/EULER/>

Comments for improvements and additions always are welcome. The project will terminate in September 2000. Then the change from a prototype to a permanent service is projected for the EULER-system. This will be supported by a consortium where some of the current partners may take part, additional partners may enter and the set of resources hopefully will be augmented.

Decomate: Unified Access to Globally Distributed Libraries

Thomas Place and Jeroen Hoppenbrouwers

Tilburg University, PO Box 90153, NL-5000 LE Tilburg, The Netherlands
{place,hoppie}@kub.nl

Abstract. The Decomate project enables mutual access to heterogeneous, distributed, and pooled digital resources of consortium members. Using a mediator architecture with a Broker and several back-end servers, a scalable and flexible system has been developed that is going in production in major European universities. Ongoing work focuses on access improvements using graphical browsing and thesaurus integration.

1 Introduction

There are several approaches to the integration of distributed, heterogeneous information sources. One approach is the implementation of architecture based on mediators [6]. In the EU funded Decomate II project [1], this architecture was used for the creation of the European Digital Library for Economics with mutual access to the heterogeneous distributed and pooled digital resources of the consortium members (the libraries of the Autonomous University of Barcelona, the European University Institute in Florence, the London School of Economics, and Tilburg University in the Netherlands). In July 2000, the project will be finished with mediator systems running at (at least) four sites in Europe giving unified access to library catalogues, article databases with links to the electronic full text, journal databases, pre-print archives, abstract and index databases, search engines indexing Web sites, and thesauri.

The content to which a mediator gives access depends on the licenses a site (a library) has acquired. Although acquiring licenses was one of the objectives of the project [1,5], this paper and the demonstration will focus on the software that was developed by the project. The software will be made available as Open Source and can be used for other disciplines than Economics; e.g., the Decomate software is used for accessing distributed image and video banks.

2 The Mediator Architecture of Decomate

In Decomate, the mediator is implemented by three sub-systems: the Broker, the Multi-Protocol Server (MPS), and the Result Optimiser (RO). The Broker, the MPS, and the RO run as separate servers that communicate with each other by using an XML-based application protocol on top of the TCP/IP communication layer. The protocol is named XREP: XML Request and Response Protocol.

2.1 The Broker

The Broker is the interface to the Decomate system for the client systems used by the end users (Web browsers or Java applets). It handles user requests, sends search and retrieval requests to the MPS, and generates XML or HTML documents to be displayed at the desktop of the user.

The Broker is designed as a XML template resolver [4]. Each request type that can be sent to the Broker has a corresponding XML template. Resolution is done by substituting XML elements by the search and retrieval results. The Broker can support several 'sites' (user interfaces). A site is a set of XML templates, organised in a hierarchy of domains and sub-domains. A (sub-)domain consists of sub-domains and XML templates. The XML templates are the leaves of a tree that represents a site.

A special site is the SiteBuilder with which other sites including itself (bootstrapping) can be built and maintained. With the SiteBuilder, XML chunks can be created that can be shared by several XML templates. Chunks defined at the level of a (sub-)domain are inherited by its sub-domains and the XML templates that belong to the domain. In this way, the Broker in combination with the SiteBuilder implements a highly customisable, flexible, and easily maintainable user interface generator.

2.2 The Multi-protocol Server (MPS)

The Multi-Protocol Server is the search and retrieval client of the distributed databases and document servers. It is controlled by the Broker that sends its XREP requests to the MPS.

Several protocols are supported by the MPS. At this moment, the MPS interoperates with Z39.50 servers and with document servers using HTTP. The MPS has an API for adding other protocols. E.g., for integrating Circulation Control functions of the OPAC of the Dutch library system vendor Pica, support of the Pica3 protocol was added to the MPS.

The MPS is a multithreaded server that allows for simultaneous (parallel) searching in databases distributed over the Internet.

Semantic interoperability is also taken care of by the MPS. The MPS is aware of the differences in query language and record syntax used by the databases that are accessed by the system. Records retrieved from the external databases are mapped to an internal element set.

2.3 The Result Optimiser (RO)

The Broker can decide to use the Result Optimiser (RO) for merging, deduplicating, and (relevance) ranking (see also section 5) result sets. For the Broker, the RO acts exactly like an MPS with added functionality. Using several 'fuzzy hash' functions [3], the RO groups records together that have a high degree of overlap, but do not need to be exactly alike. The architecture of Decomate also makes it possible to integrate advanced IR techniques with standard library databases [2].

3 Personalised Services

One of the design goals of the Decomate system was to make personalised services possible. In the present system, a Current Awareness Service and a document delivery facility is implemented.

3.1 Current Awareness

The user can ask the Broker to store queries as interest profiles. For this the Broker interacts with the Current Awareness Server (CAS), using the XREP protocol (see above). The CAS uses a relational database management system to maintain the interest profiles of the users.

The CAS Robot runs the interest profiles on a daily, weekly, or monthly basis. The frequency is determined by the user. Running an interest profile means that the CAS Robot queries the new additions to the databases. The CAS Robot uses the Broker and the MPS as mediator. The results of a run are stored in the CAS repository that is implemented as a Z39.50 database. The user is notified by e-mail of the new publications that satisfy his or her interest profile, and has access to the new results via the Broker. The results are presented to the user as issues of his or her 'personal journals.' By accessing the new issues of the personal journals via the Broker, the full functionality of the system with respect to access to the electronic full text and ordering documents is available to the user.

3.2 Document Requester

When the full text of a publication is not available on one of the document servers, the user must be given the option to order it via a document delivery service. For this the Decomate system includes a server that is called the Document Requester. The Broker can send a request to the Document Requester that is interfaced to one or more document delivery services. Which services are used depends on the library that is running the Decomate system. The Document Requester has an API that allows a library to easily add interfaces to its local document delivery services.

4 Authentication and Authorisation

For personalised services the system must know the identity of a user and whether he or she is authorised for the service. The same applies for giving access to copyrighted material (including pay-per-view access for which the user is sent a bill).

The Broker mediates between the user and the local systems for authentication and directory services (information about the users: names, addresses, email addresses, affiliations) via the Authentication broker. Different authentication mechanisms can be used in Decomate: e.g., NT authentication using Samba

and userid/password validation by a LDAP server or a dedicated password daemon. Preferably, a mechanism that allows for single-logon is used: the user logs on the network once and after that all services on the network are available to him or her. LDAP is the preferred protocol for directory services, but other implementations are supported.

5 The Concept Browser

In itself, searching via the Decomate system is rather standard. However, the project has defined a work package for the development of advanced access to federated digital libraries [3]. A special graphical tool has been developed (the Concept Browser) which allows navigation through the conceptual space of a discipline. Such a conceptual network consists of several loosely coupled vocabularies and thesauri that are implemented as Z39.50 databases using the Zthes profile for thesaurus navigation [7].

For the digital library for Economics, several vocabularies and thesauri are available, such as the commonly used JEL classification codes of the Journal of Economic Literature, and thesauri maintained by the project partners. The links between related concepts are graphically displayed and the links can be followed from concept to concept. Relevant concepts can be marked and irrelevant concepts can be crossed out. While navigating through the conceptual space, the information need of the user becomes known to the system. After the user is finished with navigating the conceptual space, the system generates an optimised query that is passed to the relevant databases. The information about relevant and irrelevant concepts is used by the Result Optimiser (see section 2.3) to rank the results.

The Concept Browser is implemented as a Java applet that interacts with the rest of the system via the Broker.

References

1. Decomate II project: <http://www.bib.uab.es/decomat2>
2. Hoppenbrouwers, J. and Paijmans, H.: Invading the Fortress: How to Besiege Reinforced Information Bunkers. In: *Proceedings of the IEEE Advances in Digital Libraries 2000 (ADL2000)*. IEEE Computer Society. 27–35.
3. Hoppenbrouwers, J.: Optimising Result Sets. In: *Proceedings of the Decomate Final Conference, June 22–23, 2000, Barcelona*.
<http://www.bib.uab.es/decomat2>
4. Kristensen, A.: Template Resolution in XML/HTML. *Computer Networks and ISDN Systems* **30** (1998) 239–249.
5. Place, T.: Developing a European Digital Library for Economics: the Decomate II Project. *Serials* **12** (2) (1999) 119–124.
6. Wiederhold, G.: Mediators in the Architecture of Future Information Systems. *IEEE Computer* **25** (3) (1992) 38–49.
7. Zthes: a Z39.50 Profile for Thesaurus Navigation.
<http://lcweb.loc.gov/z3950/agency/profiles/>

MADILIS, the Microsoft Access-Based Digital Library System

Scott Herrington, Ph.D. and Philip Konomos

Arizona State University Libraries
Herrington@asu.edu
Konomos@asu.edu

Abstract. The ASU Libraries' staff had considerable experience creating digital library systems to satisfy the needs of a major university library. These systems were designed to be high performance, large scale systems, capable of supporting very large, multimedia databases, accessible to large numbers of simultaneous users. Using this experience, the staff set out to design a digital library system that could satisfy the needs of small libraries. A small digital library system cannot be simply a scaled back version of a large system. The primary factors driving the design of a small system are cost, scalability and technical support. The resulting digital library system, named MADILIS, is designed to satisfy all of the criteria for a fully functional digital library system, while also meeting the cost, scalability and technical support needs of small libraries.

1 Introduction

The ASU Libraries' created a web-based digital library system for its locally created citation databases in 1997. This digital library system utilized a UNIX server and an unlimited user software license for BRS/Search. BRS/Search is a full-text database and information retrieval system. BRS uses a fully-inverted indexing system to store, locate, and retrieve unstructured data. The system is capable of providing rapid query response time for databases containing up to 100 million documents, each with up to 65,000 paragraphs.

The digital library system developed by the ASU Libraries included a web based search interface *and* a web-based maintenance interface, and supported links to a full range of multimedia objects. When this system was placed into full production, library systems' staff began receiving requests from other libraries for more information. Many of these requests came from small libraries without the staff to implement their own versions of this system, even when the ASU Libraries was willing to share its digital library architecture and source code.

2 Digital Library Criteria

Using a small grant from the ASU Libraries, a project was initiated to develop a digital library system capable of satisfying the needs of small and medium-sized libraries. Based on direct experience, numerous requests from other libraries and a review of library literature, the following criteria were established for this system:

1. *Low cost*, including all hardware and software components;
2. *Technically simple* to install and manage;
3. *User Friendly*;
4. *Multi-user* (including both searching and maintenance);
5. *Multimedia* digital object enabled; and
6. *Platform independent* (including both client and server components).

The project team was in unanimous agreement that every effort would be made to ensure that each of these criteria was satisfied in the design of this small-scale digital library system.

3 Design Considerations

Some design points were immediately obvious to the project team as it sought to create a system that would meet the established criteria. The client would have to be a web browser, to ensure low cost, user friendliness and platform independence. Using a standard web browser as the client also provided a solution to the issue of multi-media support: both of the major web browsers available can be used to view standard images, as well as document formats, either directly or through the installation of free plug-ins. In addition, both Windows and Macintosh operating systems include software that plays sound and video clips; these applications can be launched from standard web browsers. While this solution to supporting multi-media restricts somewhat the choice of media formats, it still leaves libraries with viable options that encompass the standard formats of choice for most library multi-media materials.

The choice of hardware to support the digital library system was based on two factors: cost and familiarity. When both of these factors were weighed, the obvious decision was to use a PC workstation as the server platform. PC workstations are the most common computers in libraries today, and—depending on configuration—are relatively affordable for most libraries. These workstations are also fairly easy to install and maintain, making technical support less of a problem.

Another fairly straightforward choice was the selection of PERL as the programming language for this project. PERL is free, readily available, platform independent and easy to learn—especially for programmers with experience in the C programming language. There is also a wealth of helpful PERL code and code modules available for download from many Internet sites.

Three critical challenges faced the designers of this new system, challenges that had not been faced in designing large scale digital library systems. These challenges were based on the need for *low cost*, *scalability* and *technical simplicity* in the new digital library system. The solutions to the problems created by these challenges were

often interrelated, and these issues were a fundamental consideration in all design decisions.

3.1 Low Cost

BRS/Search was the designers' first choice as a database management system for the small-scale digital library system, but cost led to it being ruled out as a viable option. Even considering all of the very desirable features BRS/Search provides, a multi-user license for the character-based system running on a Microsoft Windows system cost several thousand dollars. A single-user license for BRS/Search cost almost \$1,000, and would not support multiple users without considerable programming.

The search began for an alternative database management system. The designers knew that they needed a system that was low cost, scalable, multi-user, and robust. The most logical choice was the Microsoft Access database management system. This relational database management system (RDBMS) can be purchased at very low cost and is a fully functional system with a Window's client interface. It runs on inexpensive PC workstations, under both Windows 95/98 and the Windows NT operating systems. It is also a surprisingly "open" system: programs can be written that provide secure connections to Access databases created by the Windows client. Once opened, the tables in an Access database can be searched and entries modified using standard SQL commands. The Access system is inherently multi-user, and has excellent security features.

Using a relational database management system for bibliographic data does create some problems for the designers. Whereas the MARC record is designed to accommodate varying repetitions of fields from one record to the next, the RDBMS does not easily handle this concept. Creating additional fields in one record creates the same number of additional—and blank—fields in all the records in a database table, along with all the overhead this brings. This problem is solved through programming, by creating additional tables—for repeating fields—linked to, but separate from the main record in the primary table.

3.2 Scalability

Operating system upgrades often improve performance, but do not normally solve serious scalability problems. While the database management vendor also provides software upgrades, these too cannot be relied upon to solve serious performance issues. Another way to scale up a system is to move the database(s) to a more robust, higher performing database management system. Fortunately, Microsoft provides a direct path for the migration of Access databases to its enterprise level RDBMS, Microsoft SQL Server. Only minor modifications have to be made to the digital library system to accommodate this change, since the same programs—based on communicating with the RDBMS using SQL commands—will work with either RDBMS.

3.3 Technical Simplicity

Designing a system that is easy for technically naïve people to install and maintain is not a simple task. Computer software and hardware are not inherently simple to manage. Every decision made during the design of MADILIS took into consideration the need for minimal, straightforward and simple maintenance. Some of the most obvious and significant choices that affected technical simplicity included:

1. *Client software*: the system is designed to allow patrons to use any of the standard—and heavily used—web browsers that are freely available;
2. *Server hardware*: while all hardware can be difficult to support, choosing a PC workstation provides a platform that can be most easily supported by small libraries;
3. *Operating system*: finding someone to support any of the Microsoft Windows operating systems should not pose a serious problem for libraries, and the release of Windows 2000 consolidates the competing Microsoft operating systems into one;
4. *Programming language*: PERL is as accessible as any programming language available; it is free, easy to learn, and portable (with the caveat noted above); and
5. *Database management system*: Microsoft Access is inexpensive, easy to use, comes with a friendly client interface, creates accessible tables, and runs on small PC workstations.

Leveraging Electronic Content: Electronic Linking Initiatives at Arizona State University

Dennis Brunning

Arizona State University
Brunning@asu.edu

Abstract. This paper presents an overview of electronic linking initiatives at Arizona State University Libraries. It covers existing commercial solutions. These solutions include SilverLinker from SilverPlatter Information and ISILINKS from the Institute of Scientific Information. Problems, advantages, and disadvantages of these initiatives are described and explored.

1 Background

Arizona State University Libraries is a multi-campus university serving the State of Arizona, USA. ASU has a full-time enrollment of 49,000 students. The Carnegie Foundation ranks ASU as a Research I institution. The Association of Research Libraries places ASU in the top thirty of research libraries in the United States. Like many ARL libraries, Arizona State University has embarked upon an ambitious plan to create an electronic library to serve its faculty, students, staff, and community. Over the last three years, ASU Libraries has committed over 20% of its materials budget toward licensing databases and electronic journal content.

Three components make up ASU Libraries core delivery architecture. An Innovative Interfaces Innopac system provides library cataloging. A Silver Platter Electronic Resources Library (ERL) networks key indexing/abstracting services. Journal content and other information resources are mainly outsourced from library vendors and publishers. Of these outsourced services, a key indexing/abstracting service is ISI's Web of Science. All of these services are web-based. The Innopac online catalog and the ERL networked are managed locally by the library.

ASU Libraries manages and delivers electronic resources from the library's home page. Through these pages, customers have the opportunity to access over 200 indexing/abstracting services, 2000 aggregated electronic journals, and 1500 electronic journals.

2 The Problem

Most customers, including librarians, find navigating among so many electronic resources a daunting task. Despite the web environment, built around easy navigation of digital resources, our users find themselves in an information environment rich in electronic resources but poor in the area of linked relationships among resources, especially journal content.

To remedy this, and in doing so leverage our investment in web-based electronic resources, ASU Libraries has invested in the various linking initiatives which have entered production over the last year. We are also taking a keen interest in initiatives that are slated to go into production in the near future. These initiatives include SilverPlatter's SilverLinker product, ISI Isilinks, SFX linking, and soft linking from Ebscohost.

Currently, we have implemented SilverLinker and ISILINKS.

3 SilverLinker at ASU Libraries

ASU Libraries has provided SilverLinker links to electronic content since March 1999. Technically, the solution is simple. SilverLinker gathers stable URLs from publishers, creating a searchable database of link information.

We immediately saw the value of this move. Previously holdings information was limited to look-ups at the item level (journal title). Customers could (and still do!) query an alphabetical list of journals titles maintained as a web page. Item information is also linked in the InnoPac catalog using the MARC 856 tag. Finally, the ERL client, in its Windows and Web version, allows dynamic query of the ISSN field in the InnoPac item record.

With SilverLinker, users possessed linkage access to articles. In most SilverPlatter databases, a search query not only returns bibliographic information but also links to subscribed content. If the customer wants to access fulltext, he or she simply clicks on the SilverLinker icon. The customer is then linked to the publisher's web site. Linking to journal content from major indexing/abstracting services was a major step forward. Up until this point librarians were seriously concerned on how to leverage the value of indexing/abstracting information and functionality with content that resided outside these services. We were also concerned with the trend of journal publishers to segregate content at their own web sites and providing clearly less power, even inferior, search interfaces.

The fact that major publishers are cooperating with SilverPlatter (and other I/A services) to provide stable URLs to their content illustrates that a business philosophy has been revised

A particular advantage of SilverLinker is our ability to manage the SilverLinker program locally. Article level linking is necessarily a new facet of serial management. As any serials librarian knows a serials collection is a constantly changing phenomena. We weed collections, we cancel subscriptions. We select new titles. These dynamics require local choice, local decisions, and local management. Any product that doesn't allow this creates a major problem.

The biggest complaint and disadvantage of SilverLinker are probably not limited to this SilverPlatter product. SilverLinker involves the customer with two web-based products. We have found that in the hand-off of one system to another there are a number of system related problems. One is performance. Our SilverPlatter server resides in our library. The server connects directly to the campus fiber-network. This configuration enhances Internet performance on campus.

However, performance problems begin the further the user gets from our server and vary with how a customer accesses. For example, telephone access is limited to PPP modem banks that operate at very slow speeds.

This problem is exaggerated when the user is handed off from a SilverPlatter session to a session with a publisher's web site. In an instant, the customer may be a dozen or so Internet hops from the content server.

ASU does authenticate SilverPlatter users from Internet locations that do not go through the PPP modem bank. Users can access the ERL server in a peer to peer configuration. However, publisher sites vary in their ability and willingness to authenticate users who have been authenticated in SilverPlatter.

Publishers prefer proxy servers for remote authentication. For performance and management reasons, ASU Libraries has evolved a hybrid authentication system that uses referring URL. Every product on our web site has a page that provides a basic introduction to the database and ways to access the database. The customer has the choice of going through the slower modem pool or to access via ASURITE.

4 ISILINKS

ASU Libraries licensed the Web of Science early in 1968. Soon after, it became one of our more popular databases, averaging 10,000 session per month. Although the Web of Science covers arts, humanities, and the social sciences, the physical and life sciences are its major focus. ASU faculty and researchers have long had access to a local implementation of Current Contents on BRS Onsite. As the Web of Science more or less incorporated Current Contents into the Web of Science

Science, technology, and Medical publishers were among the first to create content web sites. The opportunity to bring together one of the better and more known indexing/abstracting tools and e-content was golden.

ISILINKS works very much like SilverLinker. Publishers give stable URLs to ISI who then builds a database table of valid links. These links appear in a Web of Science citation record if access is available. Some publishers have created enhanced reference links from their content. ISILINKS captures these links. So one can link back from an article to the Web of Science citation.

Unlike SilverLinker, ISI manages the linking setup. The library must inform ISI which journals to which it has electronic access. ISI has made this process quite simple with an online form submitted via email by registered contact person at the library.

At the same time, this method interferes with management at the local level. If something goes wrong with a link, the library must investigate and then communicate with ISI. And then wait for a response or a fix.

To manage effectively, the library must have a good customer support program. Feedback from customers must be verified as a problem. The problem must then be quickly and effectively communicated to ISI. If all works well, ISI fixes the problem, and communicates the fix to the library. We then test whether or not the problem has been corrected.

This process appears simple enough. Yet it isn't so smooth. ISI doesn't like to deal with just a few problems; they prefer we batch them. However, this means delay. One bad is tantamount to many bad links. The customer who can not link from

A better situation, adopted by SilverPlatter (as well as Ebsco) is self-administration. ISILINKS also does not support authentication methods useful to ASU Libraries. They support IP filtering or proxy servers. As noted, ASU Libraries have decided that proxy service does not meet performance requirements. Moreover, we can not manage a sufficiently robust proxy service, which would meet performance requirements.

5 What We Have Learned and Other Observations

5.1 Leveraging

The basic complaint against publisher web sites lodged by librarians has been the lethargic and function-handicapped search engines provided by the publishers. System librarians can add to this complaint that publishers, new to online services, do not seem to act like online services. Customer support is rarely 24X7 and publisher web sites do not seem to take seriously the need to communicate downtimes to users.

In complete contrast, indexing/abstracting services have a long history of working as online services. Support is 24X7 and customer support is well established. Search software has evolved over the last twenty to forty years to very effective and powerful retrieval tools.

5.2 Eliminating Intermediate User Steps

Librarians and customers alike also complain about the steps required to move from a finding tool to a document location. Hosting services that aggregate electronic content have long succeeded by merging the finding tool with the actual content, e.g., Bell and Howell Proquest, Ebscohost. Linking initiatives like SilverLinker and ISILINKS emulate these aggregators by bridging powerful finding tools and content.

5.3 Maximizing the Web Model

Web use is click-conscious and click-oriented. Users expect to be linked from one piece of information to another piece or related or more useful information. Linking initiatives play into this model.

5.4 Problems with Standards

No standards exist for linking. The CrossRef initiative promises to create a system of providing linking information from publishers to a clearinghouse for such information. The Publishers International Linking Association will manage this clearinghouse (see <http://www.crossref.org/>).

At present, SilverLinker and ISILINKS support proprietary solutions. There is no reason for either of these products to adhere to a proprietary solution should a standardized approach be developed.

5.5 Authentication

Authentication presents a stickier problem. Most publishers adhere to IP domain restriction for academic customers. This approach is easy to implement, maintain, and offers a great degree of security. Unfortunately, it presents specific problems for linking initiatives.

On one hand, if the indexing/abstracting service and publisher are both IP authenticated, then a user accessing from a valid IP range will have no problem moving from one IP restricted service to another. He or she will be validated by the service accessed.

On the other hand, if a valid user accesses from a restricted domain, the library must take certain steps to authenticate this user. Generally speaking this involves some form of query to a database of valid users and passing to the information provider some piece of information that says this user may access a set of resources.

On the surface this appears to be a simple process. In reality, there are many obstacles. Many publishers protect their content by limiting authentication options. For large institutions publishers prefer IP restriction or issuing userid and passwords. These encryptions are changed frequently to maintain security. As a result, the library inherits a considerable management problem of determining who should receive

userid's and passwords, updating lost and forgotten passwords, issuing revised passwords etc.

The main problem for linking involves the continuity of authentication across what may be called authentication boundaries. A user needs to be taken from one service to another and back—as many times as required. Current authentication policies and technologies do not readily support this.

5.6 Performance Problems

Best performance over Internet bandwidth occurs in situations where the user is close to a server. Theoretically, the Internet forms an ideal distributed environment where servers may reside anywhere in the world.

In fact, the Internet does not constitute an ideal distributed environment. Servers more proximate to each other on the Internet tend to perform better as distributed servers than do servers that are located far from each other.

Linking of services encounters real problems of performance as a result of server locations and server capacities. Presently, publisher servers do appear to be scaled to perform well and efficiently for all users.

5.7 Lost in Cyberspace?

At present, over 200 information providers comprise ASU Libraries set of web services. Over 3,000 electronic journal titles are available from an array of aggregators and publishers. To provide sensible and simple navigation among these resources stands as major challenge. Linking through SilverPlatter and ISILINKS accomplishes the important task of bringing together indexing resources and content. The remaining task will be to make easier the customers understanding the various informational “worlds” they enter when they click on links.

Asian Film Connection: Developing a Scholarly Multilingual Digital Library – A Case Study

Marianne Afifi

Center for Scholarly Technology, Information Services Division
University of Southern California
Los Angeles, CA 90089-0182
afifi@usc.edu

Background

In 1998, the staff of the Center for Scholarly Technology (CST) in the Information Services Division (ISD) at the University of Southern California (USC) was approached for assistance with a database/digital library project. Generally, the role of the Center is to assist with curricular technology projects. Although the project described below is somewhat outside the normal scope for the Center, the scholarly nature of the project was taken into account in granting assistance to it. The impetus for the project came from Jeanette Paulson Hereniko, Director of the Asia Pacific Media Center at the Annenberg Center for Communication and Founding Director of the Hawaii International Film Festival. Ms. Paulson organized a conference in Los Angeles in May of 1998 attended by members of NETPAC, the Network for the Promotion of Asian Cinema, a pan-Asian cultural organization involving critics, filmmakers, festival organizers and curators, distributors and exhibitors, and film educators. At the conference, a plan for the creation of a scholarly, multilingual digital library about film in Asia was first presented.

Karen Howell, the Director of CST, working closely with Ms. Hereniko, developed a planning document for the conference. Other CST team members who contributed during the course of the project are Marianne Afifi, Robert Doiel, Dan Heller, Apryl Lundsten, and Bonnie Ko, an intern from UCLA. For the construction of the promotional website, we also were fortunate to have the help of Dr. Shao-yi Sun who will be managing the project as more funds become available.

The purpose of the digital library is to promote the marketing of as well as awareness and education about film in Asia, not only to the rest of the world, but also among the Asian countries themselves. Asian films are rarely seen outside their respective countries and scholarly information about them is sparse and often not available in other countries. Thus the Internet can serve as a vehicle for communication and dissemination of information that has so far been difficult to collocate. The digital library's audience is expected to consist of filmmakers, critics and journalists, distributors and programmers, scholars and students, cinema fans, and anyone interested in Asian cultures. The participants in the conference were enthusiastic about the project and accepted the responsibility for conversion of their respective data to digital format and for contributing the data to the digital library.

Plan for a Digital Library

The plan was to include initially five countries that would be contributing content to the database, China, Japan, India, Korea and Taiwan. It was anticipated that, once a prototype system was established, other countries, primarily in Asia would be approached to participate. Content was defined as digital objects such as text, video files such as film clips, still images, audio files such as soundtracks and other digital objects to be defined in the future. For all released films, 15 basic fields were initially defined, with many fields having two and up to 11 subfields. Each country was also to select 8-15 films from 1995 through 1997 to be highlighted. Additional fields would be added for the highlighted films. To be highlighted, a film needed to meet certain agreed upon criteria.

A central database was to be set up to facilitate the input of data and metadata in each country. Except for India, where English is the national language, it was expected that the input interface would be in the language and script of the country doing the data input and cataloging. It is important to note here that China and Taiwan use slightly different character sets, those in China having been simplified over the years. Thus we expected to be dealing with five character sets.

The central database was to be kept in the United States, initially at USC. Here, the data would be translated into the remaining languages, the translations entered and the integrity of the data that was input elsewhere verified. In addition, all input and output interfaces would be created centrally and a database manager would be hired to manage the flow of data. In addition, USC was to handle database administration on a central server. Due to lack of sufficient funds for the purchase of robust database software, it was anticipated that database software currently available at USC was going to be used in the project

The project plan was developed using a matrix that listed tasks on one axis and the corresponding responsibility, costs, funding sources, timeline, and review on the other. The sequence of tasks began with the initial fundraising, followed by setting common conventions, data gathering, translation, database design and construction, website creation, website update, and promotion and publicity about the site. So far, some of these tasks have been accomplished in part.

Collaboration with the Center for Software Engineering

To move the project forward, the CST team collaborated with the Center for Software Engineering at USC. Through a collaborative venture of that Center and the ISD, every year the students in the Computer Science 577 course take on software development projects that are proposed by staff in the ISD. These projects are often too time consuming or too technologically challenging to be developed or implemented by the staff. Students in the class help to get the projects underway by playing the roles of software engineers and creating a system design or plan for the “clients”.

We proposed the Asian film database project as one of the student projects; specifically we wanted the students to perform a systems analysis, define system and software requirements, design a system and software architecture, as well as study the feasibility of the project. A group of five students accepted to work on all these tasks for the project. The CST team met with the students on a weekly basis and worked with them to develop a design for the digital library. This project occurred in the fall semester of 1998 and many of the recommendations are still valid despite the rapid progress of information technology.

Promotional Website and Prototype Phase

During the time at which fundraising took place, a promotional website was created to show the potential of what the digital library could look like and also to attract funding for the project. The site does not contain a database but consists of web pages simulating a database-driven information architecture. The website will be further developed to serve as a template for a prototype for the Asian Film Connection database that is expected to be created later in the year 2000 using content from Japan. The CST has also been considering the use of open source applications for this prototype in order to ease the burden of funding.

Funding

Funding for this project has been difficult despite the cooperation and good intentions of the contributing countries. The funds for the initial Annenberg Center grant were limited to three years. The intellectual work of CST and the infrastructure support from the Information Services Division, such as the provision of equipment, networking and disk space, has been an in-kind contribution and continues to be so until now. The work of the CS577 course created a useful vehicle for the students to do their work, but was also free of charge. Ms. Hereniko has spent much of her time on fundraising efforts, which she continues, so that the Asian Film Connections digital library may indeed become a reality.

Social Aspects

Although, on a management level, country representatives agreed that this digital library was a good idea, the reality of developing such a product was marked by communication problems on several levels. Only some of the designated contacts responded to calls for materials. It was difficult to ascertain what kind of technology infrastructure could be assumed at the different sites. Another initial difficulty was how to communicate the concept of a database or digital library to some of the participants and also to funders without a prototype. Once the promotional website was developed, it was easier to approach funders because the nature of the digital library could be explained more easily. We also could show different character sets, images, and other objects.

Despite the cooperative nature of the initial conference, it was clear that we had to use care not to offend any country. Although we did not necessarily think that the English version should show up first, we decided to do so, because the management of the site would be in an English-speaking country. Because of the political situation between China and Taiwan, we had to be careful not to favor one over the other, for example in listing the countries. In planning the translations, we have to be careful to have all translations done before we upload them into the database. Because there were so many groups involved at USC, the internal communication and each contributor had different expectations of the project, which sometimes led to misunderstandings.

Conclusions

This project teaches us several lessons. One is that despite the proliferation of commercial sites and databases about film, scholarly sites that aim to be objective and not influenced by commercial interests are difficult to establish. In this case the representatives of the Asian film industries were very willing to contribute their time but they did not have additional resources to commit to the project perhaps due to an economic downturn in Asia at the time.

Another lesson to be learned is that multilingual digital libraries are exciting and promise to improve communication among countries without their having to give up their linguistic identities in favor of English which some say is the new Esperanto. Films from Asia and their directors are virtually unknown in the West and digital libraries about them may increase international recognition of their productions. Although much progress has been made with the representation of different scripts on the Internet, there are still costs associated with translation. During the course of the project we calculated that the cost for translation was a big part of the budget.

Furthermore, it appears that funding for scholarly endeavors from foundations and other traditional funders of educational projects is difficult to obtain in a "dot com" world. Whereas startups and corporations pay sizeable sums for the design of often marginally effective web sites, funders of educational web sites were not willing to underwrite the costs of this project, although the costs were realistic relative to the project goals.

Useful URLs

Asian Connections: <http://www.asianfilms.org>
Center for Scholarly Technology: <http://www.usc.edu/cst>
Information Services Division: <http://www.usc.edu/isd>
University of Southern California: <http://www.usc.edu>
Center for Software Engineering at USC: <http://sunset.usc.edu/>

Conceptual Model of Children's Electronic Textbook

Norshuhada Shiratuddin¹ and Monica Landoni²

¹ School of Information Technology, Universiti Utara Malaysia,
shuhada@dis.strath.ac.uk, shuhada@uum.edu.my

² Department of Information Science, University of Strathclyde,
monica@dis.strath.ac.uk

Abstract. First step in developing electronic book is to build a conceptual model. The model described in this paper is designed by integrating Multiple Intelligences Theory with existing electronic book models. Emphasis is on integrating the content of a page with appropriate activities that meet and cater for the diversity of learning styles and intelligence in young children. We postulate that an additional feature for children e-book would be to present contents by mixing different presentation modes and including various activities which support as many intelligences as possible.

1 Introduction

In most software development process, the first and foremost stage is to build ideas and present them in the most effective ways. This is done through building a concept, often referred to as concept development phase. It involves the task of inventing and evaluating idea [1] and is more apparent in the development of interactive multimedia programs. Since electronic book (e-book) is an example of interactive programs, its development has also to go through this phase.

We are currently planning to develop a children multimedia electronic textbook (children ages in between 5-9 years old) which support and cater for seven different intelligences and learning styles. These seven intelligences were proposed by Howard Gardner [2] in his Multiple Intelligences theory. This theory states that there exist at least seven intelligences (thus seven learning styles): verbal/linguistic, logical-mathematical, visual/spatial, bodily-kinesthetic, musical, interpersonal and intrapersonal:

- A child is verbal/linguistic if he loves words and enjoys reading, writing and story telling.
- A logical-mathematical child is more interested in concepts, numbers and scientific exploration.
- A visual/spatial child learns best through pictures and images, enjoys art and mentally visualizes things easily.
- A bodily-kinesthetic child needs to move and touch to learn.
- A musical child uses rhythm and melody to learn.
- Interpersonal child learns best with other people around.
- Intrapersonal child gets more out of being left alone to learn.

Armstrong [3] identifies that each child could learn in any one of these ways or through a combination of several ways. We believe that e-books could play a crucial role by supporting this learning process. In order to proceed with the e-book development, the conceptual model has first to be built. This paper presents and discusses our proposed model.

2 E-Book Models

Formal models by Barker & Manji [4], Barker & Giller [5], Stephen et. al. [6], and Barker [7] of e-books already exist which describe the structure of e-books. Functional aspects and its definitions are also available and described in details by Catenazzi [8] and Landoni [9]. However, none of these models discussed about the presentation of content and activity involved in the book pages. Because we are developing children multimedia electronic textbook, a great emphasis on the content of each page and the activities involved in making sure interaction and learning occur have to be taken.

The following section describes the conceptual model of our proposed e-book. This model is defined in terms of structure and presentation. It is built to cater for the seven intelligences in line with Gardner's theory.

2.1 Children's E-Book Conceptual Model

2.1.1 Structural Components

Before the structure of our children e-book can be explained, a comparison of the printed version of children and adult textbooks has to be performed. This needs to be done since as mentioned earlier detailed e-book models do already exist. However, they are developed and implemented in different environment. The existing models are generalized models more suited to higher-level textbooks and scientific books. Our book on the other hand is designed for young children. Furthermore, designing e-book based on printed book metaphor has proved to be helpful in lessening the cognitive load of end users [10].

A book is made up of at least three main sections: front section, main section, and back section. Each of this section is further made up of subsections. The following table is comparison based on a sample of 5 children textbooks and 5 adult textbooks with various subjects.

By adopting mathematical sets and with reference to the above table, it can be concluded that the structural components of children multimedia e-book are:

```

Book = {Front Section, Main Section, Back Section}
Front Section = {Title page, Verso page, Table of Content}
Main Section = {Chapter i} for  $\forall i \in N$ ,
    where  $N = [1, 2, 3, \dots]$ 
    Chapter i = {Page i} for  $\forall i \in N$ 
    Page i = {Header, Paragraph j} for  $j \leq 5$  and  $\forall i \in N$ 
    Paragraph j = {Text, Graphics} for  $j \leq 5$ 
Back Section = {Back cover}

```

Table 1. Comparison between adult and children printed textbooks

Section	Structure	Most adult printed books	Most children printed books
Front	Title page	Yes	Yes
	Verso page	Yes	Yes
	Abstract	Yes	No
	Foreword	Yes	No
	Preface	Yes	No
	Acknowledgement	Yes	No
	Dedication	Yes	No
	Table of Content	Yes	Yes
	List of tables	Yes	No
	List of figures	Yes	No
Main	Chapters	Yes	Yes
	Pages	Yes	Yes
	Header	Yes	Yes
	Paragraphs	Yes	Yes
	Text	Yes	Yes
	Graphics	Yes	Yes
	Tables	Yes	No
	Figures	Yes	No
	Links	Yes	No
	Footnote	Yes	No
Back	Back Cover	Yes	Yes
	References	Yes	No
	Index	Yes	No
	Glossary	Yes	No
	Related documents	Yes	No
	Biographical	Yes	No
	Appendix	Yes	No

In the above definitions, main section can have as many chapters as required and in each chapter there are no limit to the number of required pages. However, in each page, the number of paragraphs for children textbooks is frequently less than five, with an average of three.

2.1.2 Content Presentation Components

Designing e-book for learning needs greater effort in the presentation of the book contents. It is important to carefully design the way content is structured, organized, and presented. The types of activity in which the users will be involved play significant roles in the success of pedagogic design [4]. Thus, studies on what kind of activities [11] cater for most users' needs are indeed helpful in promoting better children e-book design.

Content presentation should be viewed in terms of 3 criteria [12] which are the number of separate intelligences it engages, the extent that each is engaged and how well content can be accessed through each intelligence. With regard to these criteria, listed below are the appropriate activities, which should be included in the design of children e-book contents so as to meet the seven intelligences:

1. Activities which should be considered when meeting *verbal/linguistic* intelligence are writing essays, poems, articles, short play with word processors, annotating voice, recording speech, reading, story telling, debating practice programs, interviewing with programs, explaining article and listening.
2. Activities that can be taken into account when meeting *visual/spatial* intelligence are drawing programs, painting programs, using spreadsheet to create charts and diagrams, interacting with interactive maps, manipulating digital images, taking digital photographs, building 2D and 3D models, and interacting with animation or motion pictures.
3. Activities which should be included when catering *musical* intelligence are story telling with songs, chanting, sing along programs, creating songs, making musical instrument, and listening to music, rhythms and rhyme.
4. Activities that should be included when designing for *logical-mathematical* intelligence are playing electronic games, puzzles, strategic games and logic games, calculating and mathematics programs, making estimation, predicting story, working with geometric shapes and patterns, and solving mysteries or problems.
5. Activities that match *bodily-kinesthetic* intelligence are inputting data using alternate input such as joy stick, mouse, and touch screen, allowing users to move objects around the computer screen, making a lot of eyes movement with animation, providing eye-hand coordination games, asking users to dance and act, and providing hands-on construction kits that interface with the computer.
6. Activities which should be considered when meeting *interpersonal* intelligence are using email and chatting programs, allowing games which require two or more players, providing instruction for group activities, desk top conferencing and meeting and listening to other user on-line.
7. Examples of activities that cater for *intrapersonal* intelligence include providing drill and practice programs, playing games which the opponent is the computer, creating notes on daily activities/on-line diary and assessing user own work.

The content presentation components of children e-book should include these activities. We propose four different presentation modes in each page. Content in any page is presented by using four objects and these objects contain programs with activities that support the seven intelligences. In order to do this, the definition of Page $i = \{\text{Header}, \text{Paragraph } j\}$ in the previous section needs to be changed to:

Page $i = \{\text{Contents}, \text{Objects}\}$ for $\forall i \in N$, where $N = [1, 2, 3, \dots]$
 Contents = $\{\text{Header}, \text{Paragraph } j\}$ for $j \leq 5$
 Paragraph $j = \{\text{Text}, \text{Graphics}\}$ for $j \leq 5$
 Objects = $\{\text{Object A}, \text{Object B}, \text{Object C}, \text{Object D}\}$
 Object A = $\{\text{Graphic page}, \text{Program } G_i\}$ for $\forall i \in N$
 Object B = $\{\text{Talking page}, \text{Program } T_i\}$ for $\forall i \in N$
 Object C = $\{\text{Hypermedia page}, \text{Program } H_i\}$
 for $\forall i \in N$
 Object D = $\{\text{Web page}, \text{Program } W_i\}$ for $\forall i \in N$

Remark 1. Program refers to the kind of technology appropriate for an activity, G = graphic, T = talking, H = hypermedia, and W = web.

The above definition describes each page as containing contents plus a combination of any of four different objects. The content will consist of the header

and the paragraphs, which usually are made of text and graphics. The contents are also supported and reinforced by using graphic, talking, hypermedia and/or web pages. And in each of these four different types of pages, a collection of appropriate activities that match the seven intelligences will also be included. Activity is presented in the form of program. The higher the number of i , in the definition Program $G_i / T_{ij} H_{ij} W_i$ the more activities are provided in the e-book.

3 Conclusion

In conclusion, we postulate that an additional feature for children e-book would be to present contents by mixing different presentation modes and including various activities which support as many intelligences as possible. This is an on going project, and in order to prove the above assumption, the conceptual model described in this paper will be investigated in the near future. Investigation will be performed to evaluate users' satisfaction on the concept of children e-book that matches activity with intelligence.

References

1. Stansberry, D.: Labyrinths: the art of interactive writing and design:content development for new media, Wadsworth Pub., ITP (1998)
2. Gardner, H.:*Frames of Mind: The Theory of Multiple Intelligences*, Fontana (1993)
3. Armstrong, T.:*Multiple Intelligences in the Classroom*, Assoc. for Supervision and Curriculum Development, Alexandria, USA (1994)
4. Barker, P.G and Manji, K.:Designing Electronic Books, Educational and Training Technology International (1991) 28(4) 273-280
5. Barker P.G and Giller, S. An Electronic Book for Early Learners, Technology International (1991) 28(4), 281-290
6. Stephen, R., Barker, P., Giller, S, Lamont, C., and Manji, K.: Page Structures for Electronic Books, Educational and Training Technology International (1991) 28(4) 291-301
7. Barker, P.G.: Electronic Libraries of the future, *Encyclopedia of Microcomputers*, (1999) 23(2) 121-152
8. Catenazzi, N.: A Study into Electronic Book Design and Production: Hyper-Book and the Hyper-Book Builder, PhD thesis, Uni. Of Strathclyde (1993)
9. Landoni, M.: The Visual Book System: A Study of the Use of Visual Rhetoric in the Design of Electronic Books, PhD thesis, Uni. Of Strathclyde (1997)
10. Landoni, M., Crestani, F., and Melucci, M.: The Visual Book and the Hyper-Textbook: Two Electronic Books One Lesson?, RIAO Conference Proceedings, (2000) 247-265
11. Pickering, J.C.: Multiple Intelligence and Technology: A Winning Combination, Teachers in Technology Initiative, The University of Rhode Island & Rhode Island Foundation (1999)
12. Fetherston, T.: A socio-cognitive framework for researching learning with interactive multimedia, Australian Journal of Educational Technology (1998) 14(2) 98-106

An Information Food Chain for Advanced Applications on the WWW

Stefan Decker¹, Jan Jannink¹, Sergey Melnik¹, Prasenjit Mitra¹, Steffen Staab², Rudi Studer², and Gio Wiederhold¹

¹ Computer Science, Stanford University, Stanford, CA, 94305, U.S.A.

² University of Karlsruhe, Institut AIFB, 76128 Karlsruhe, Germany

{stefan, jan, melnik, prasen9, gio}@db.stanford.edu

{staab, studer}@aifb.uni-karlsruhe.de

Abstract. The growth of the WWW has resulted in amounts of information beyond what is suitable for human consumption. Automated information processing agents are needed. However, with the current technology it is difficult and expensive to build automated agents. To facilitate automated agents on the web we present an *information food chain* for advanced applications on the WWW. Every part of the food chain provides information that enables the existence of the next part.

1 Introduction

The growth of the World Wide Web has resulted in large amounts of information available for human consumption. Since humans have a limited capacity for processing information, we need automated information processing agents [GK94]. However, with the current technology it is difficult and expensive to build automated agents, because agents are not able to understand the meaning of the natural language terms found on today's webpages. To facilitate automated agents on the web, agent interpretable data is required. Creating and deploying data about a particular domain is a high effort task, and it is not immediately clear how to support this task. This paper presents an *information food chain* [E97] for advanced applications on the WWW. Every part of the food chain provides information that enables the existence of the next part.

2 The Information Food Chain

For data exchange on the web, it is necessary to have a specification of the terminology of the domain of interest. *Ontologies* [FH97] are a means for knowledge sharing and reuse, and capture the semantics of a domain of interest. An ontology is a formal specification of vocabularies used to describe a specific domain. It provides a basis for a community of interest for information exchange. Since there will be no ontology available describing all possible domains, multiple ontologies for different application domains are needed. Ontologies have been defined using

RDF [LS99], RDF Schema [BG99], XOL, [KCT99]. An XML-based representation language and formal ontologies are a foundation for automated agents on the web. However, to create machine-interpretable data and deploy it, we need infrastructure (e.g. support- and deployment tools), given by an information food chain (see Fig. 1).

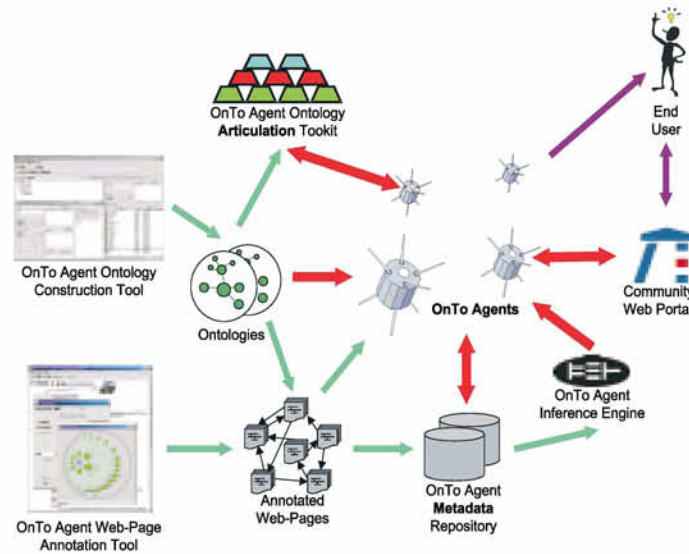


Fig. 1. Agents Information Food Chain

Ontology Construction Tool The food chain starts with the construction of an ontology. Constructing and maintaining an ontology involves human interaction. Also ontologies evolve and change over time (as our knowledge, needs, and capabilities change) so reducing acquisition and maintenance cost of ontologies is an important task. An *Ontology Construction Tool* is necessary to provide the means to construct ontologies in a cost-effective manner. Examples of ontology editors are e.g. the Protégé framework for knowledge-base system construction¹, and the WebOnto-Framework [D98]. WebOnto supports collaborative, distributed browsing, creation and editing of ontologies by providing a direct manipulation interface that displays ontological expressions. However, none of those tool is yet used for creating ontologies for web based agent applications.

Webpage Annotation Tool For information on webpages to be machine-interpretable, semantic information about the content of the page is needed. A *Webpage Annotation Tool* provides an annotator the means to browse an ontology and to select appropriate terms of the ontology and map them to sections of a webpage.

¹ <http://smi-web.stanford.edu/projects/protege/>

Using Ontopad [DEF99], an enhanced HTML editor, the annotator can select a portion of the text from a webpage and choose to add a semantic annotation, which is inserted into the HTML source. However, for significant annotation tasks a practical tool also has to exploit information extraction techniques for semi-automatic metadata creation. Often it is sufficient to give the user a choice to annotate a phrase with some ontological expressions. Resources like WordNet [F98] and results obtained from the Scalable Knowledge Composition (SKC) project [MGK00] can be used to annotate webpages semi-automatically.

Ontology-Articulation Toolkit In order to solve a task that involves consulting multiple information sources that have unknown ontologies, an automated agent needs to bridge the semantic gap between the known and the unknown ontologies. In the Scalable Knowledge Composition Project (SKC [MGK00]) we have developed tools that automatically generate articulations or semantic bridges, among multiple ontologies (by consulting online dictionaries and using other heuristics), and presents it to an expert for validation.

Agents Inference System For declarative information processing an agent needs an *Inference Engine* for the evaluation of rules and queries. An inference engine helps to exploit available metadata by inferring further implicit metadata [DBS98]. The properties of the reasoning capabilities have to be carefully chosen. A reasoning mechanism, which is too powerful, has intractable computational properties, whereas a too limited approach does not enable the full potential of inference. Deductive database techniques have proven to be a good compromise between these tradeoffs.

Automated Community Portal Site Of course, the annotation process itself has a human component: although the effort for generating the annotation of a webpage is an order of a magnitude lower than the creation of the webpage itself, there has to be some incentive to spend the extra effort. The incentive for the creation of the annotation (which is metadata for the webpage) is visibility on the web, e.g. for a *Community Web Portal*, which presents a community of interest (distributed on the web) to the outside world in a concise manner. The data collected from the annotated webpages helps automate the task of maintaining a Community Web Portal drastically. A Semantic Community Portal Website is an application demonstrating the use of ontology-based markup.

3 Related Work and Conclusion

In the Ontobroker [DEF99] and SHOE project [HHL98] means were investigated to annotate webpages with ontology-based metadata. - thus realizing part of the food chain. Agent based architectures usually focus on inter- agent communication instead to ontology creation and deployment (see [NN99] for a complaint about neglecting of the ontology problem).

We presented an information food chain, that empowers intelligent agents on the web and deploys applications that will facilitate automation of information

processing on the web. Fundamental to that approach is the use of a formal markup language for annotation of web resources.

We expect this information infrastructure to be the basis for the "Semantic Web" idea - that the World-Wide-Web (WWW) will become more than a collection of linked HTML-pages for human perusal, but will be the source of formal knowledge that can be exploited by automated agents. Without automation, and precision of operations, business and governmental uses of the information will remain limited. This food chain is partially implemented. It will be completed in the Onto-Agents-Project at Stanford University - funded in the DAML program of the DARPA.

References

- [BG99] D. Brickley, R. Guha: Resource Description Framework (RDF) Schema Specification W3C Proposed Recommendation 03 March 1999, <http://www.w3.org/TR/PR-rdf-schema/>
- [DBS98] S. Decker, D. Brickley, J. Saarela, and J. Angele: A Query and Inference Service for RDF. In: Proceedings of the W3C Query Languages Workshop (QL'98), <http://www.w3.org/TandS/QL/QL98/pp.html>, 1998.
- [DEF99] S. Decker, M. Erdmann, D. Fensel, and Rudi Studer: Ontobroker: Ontology Based Access to Distributed and Semi-Structured Unformation. In: IFIP TC2/WG2.6 Eighth Working Conference on Database Semantics (DS-8), Kluwer, 351-369, 1999.
- [D98] J. Domingue: Tadzebao and webonto: Discussing, browsing, and editing ontologies on the web. In: Proc. of KAW98, Banff, Canada, 1998, <http://ksi.cpsc.ucalgary.ca/KAW/KAW98/domingue/>
- [E97] O. Etzioni: Moving Up the Information Food Chain: Deploying Softbots on the World Wide Web, *AI Magazine*, 18(2): Spring 1997, 11-18.
- [F98] Christiane Fellbaum (ed): Wordnet: An Electronic Lexical Database; MIT Press, 1998.
- [FH97] N. Fridman Noy and C. D. Hafner: The State of the Art in Ontology Design, *AI Magazine*, 18(3):53-74, 1997.
- [GK94] M. R. Genesereth, S.P. Ketchpel: Software Agents. In: *Communications of the ACM* 37 (7, July), 48-53, 1994.
- [HHL98] J. Heflin, J. Hendler, and S. Luke: Reading Between the Lines: Using SHOE to Discover Implicit Knowledge from the Web. In: *AAAI-98 Workshop on AI and Information Integration*, 1998.
- [KCT99] Peter D. Karp, Vinay K. Chaudhri, and Jerome Thomere: XOL: An XML-Based Ontology Exchange Language. <ftp://smi.stanford.edu/pub/bio-ontology/xol.doc>
- [LS99] O. Lassila, Ralph Swick: Resource Description Framework (RDF) Model and Syntax Specification, W3C Recommendation 22 February 1999, <http://www.w3.org/TR/REC-rdf-syntax/>
- [MGK00] P. Mitra, G. Wiederhold, and M. Kersten. A Graph-Oriented Model for Articulation of Ontology Interdependencies. In: *Proceedings Conference on Extending Database Technology 2000 (EDBT'2000)*, Konstanz, Germany, 2000.
- [NN99] H. S. Nwana, D. T. Ndumu, A Perspective on Software Agents Research, In: *The Knowledge Engineering Review*, Vol 14, No 2, pp 1-18, 1999.

An Architecture for a Multi Criteria Exploration of a Documents Set

Patricia Dzeakou¹ and Jean-Claude Derniame²

¹IRD, Institut de Recherche pour le Développement, 5 R. du Carbone 45072 Orléans France
Patricia.Dzeakou@orleans.ird.fr

²LORIA, Laboratoire Lorrain de Recherche en Informatique et ses Applications, Campus
Scientifique BP 239 - F54506 Vandœuvre-lès-Nancy France
Jean-Claude.Derniame@loria.fr

Abstract. This paper presents an architecture suitable for exploring a set of documents depending on multi criteria documents. During the exploration session, the user progressively builds a portfolio of relevant documents using semantic views.

Introduction

Collecting documents for a study is a long process [1]. The documents collected are multimedia documents described using metadata expressing their multiple characteristics. We propose an architecture for a storage and retrieval system dedicated to the multi criteria exploration of a document set. We use a particular domain to identify the requirements and to validate our architecture. Our approach allows to gradually explore a set of such documents with many retrieval paths.

Architecture

The core elements of the architecture framework [2] are the Document Repository, the Metadata Repository and the selection criteria. The selection criteria are integrated in the document metadata model as document descriptors. This model serves to generate the form used by the document providers. To allow a gradual exploration, we introduce two concepts: the portfolio and the view on the portfolio.

1. A portfolio is a kind of dynamically evolving container of selected documents which is persistent in the framework. Applying selection criteria on a portfolio builds a new one (basic operation of the exploration process).
2. A view is a set of selection criteria on a portfolio, corresponding to a specific aspect. It is characterised by a specific user interface with the following components: 1) a navigation context window where are displayed current information about the exploration, 2) a navigation map which provides means to explore the document set using the view's selection criteria [3] and 3) a portfolio content list. Therefore, the view offers a gateway for building new portfolios. Four

views templates have been designed. They differ by the provided selection criteria in the navigation map: 1) a view template using descriptors coming from the metadata model, 2) a view template using a fragment of an taxonomy, 3) a view template using a map and 4) a view template using predefined usage metadata [4].

Overview of the architecture

The proposed architecture can be implemented on an Intranet with three services: 1) the Gatherer Service allowing registered users to add documents to the Document Repository and to provide metadata, 2) the User Interface Service managing display of views according to user's actions, and 3) the Administration Service giving access to adaptation functionalities such as creating and specifying views and parameters. These services rely on three knowledge warehouses as described in figure 1.

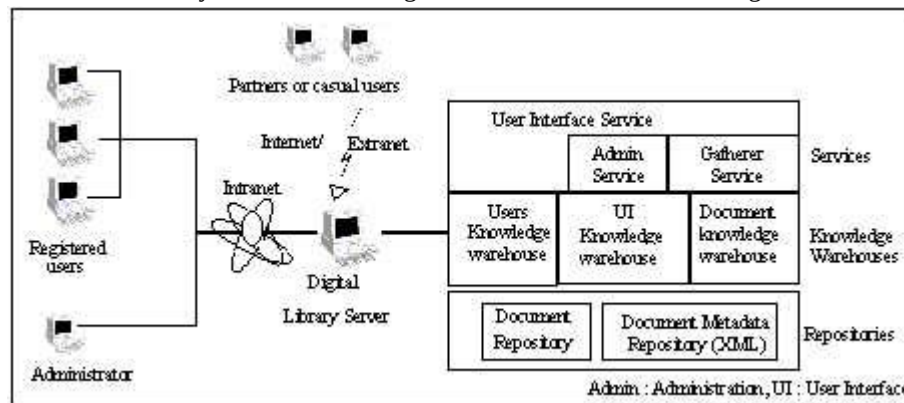


Fig. 1. Overview of the architecture

Application on an Environmental Document Set

To illustrate our conceptual view of the system, as well as to demonstrate its feasibility, here is described a scenario of an exploration process done on our prototype with images, articles about the delta of Niger. Four views have been created by the administrator: 1) a publication view with tools for selecting documents on the basis of criteria like author or document type, 2) a thematic view with tools for selecting documents using as criteria the elements of a taxonomy fragment, 3) a geographic view with tools for selecting zones and 4) a gregarious view with tools for selecting documents using criteria like group repartition, document consultation rate.



Fig. 2. In this thematic view elements relative to production systems are shown. The user has selected "pêche". The current portfolio length and the historic of queries are modified



Fig. 3. A geographic view showing a basic layout of the delta of Niger

Assuming the user is interested in photographs about fishing taken in a region. The user can first use the thematic view to retrieve documents on fishing (fig. 2). Then he can switch to a geographic view (fig. 3) showing a geographic map divided into

zones. By selecting the desired zone, he restricts his portfolio to documents referring to both fishing and the selected zone. To keep only photographs in his portfolio, he can select the criteria "document type" in the publication view. Due to the portfolio size reduction, the speed of view's display and retrieval is enhanced [3]. An operator is also provided allowing to browse the portfolio content in another window (fig. 3).

Conclusion and Perspectives

The efforts towards a global digital library have focused on the interoperability of existing document databases [5] [6]. However, our approach put the accent on digital libraries gradually built within an institution [7]. This paper presents a framework architecture for a document storage and retrieval system suitable for complex multi criteria selection process. Is also shown an instance of this framework applied to environment (<http://www.orleans.ird.fr/WISEDL>). This system will be integrated on the SIMES/WISE-DEV¹ project platform currently under progress.

References

1. Dzeakou, P., Morand, P., Mullon, C., Méthodes et architectures des systèmes d'information pour l'environnement. Proceedings of CARI'98, 1998
2. Ron, D., Jr, Lagoze, C., Payette, S., A Metadata architecture for Digital Libraries, Advances in Digital Libraries, Santa Barbara, 1998
3. Plaisant, C., Schneiderman, B., Doan, K., Bruns, T, Interface and Data Architecture for Query Preview in Networked Information Systems, ACM Transactions on Information Systems, Volume 17, N°3, 1999, 320-341
4. Zhao, D., G., Ramsden, A., The ELINOR Electronic Library, Digital Libraries - Research and Technology Advances, ADL'95 Forum, 1995, 243-258
5. Leiner, B., The NCSTRL Approach to Open Architecture for the Confederated Digital Library, D-Lib Magazine, 1998
6. Nebert, D., Information Architecture of a Clearinghouse, WWW Access to Earth Observation and Geo-Referenced Data Workshop, 1996
7. Staab, S., Decker, S., Erdmann, M., Hotho, A., Mädche, A., Schnurr, H. P., Studer, R., Semantic Community Web Portals, submitted to WWW'9. Amsterdam 2000.

¹ Scientific project's partners are Western and Subsaharian institutions and the financial parties are CEE and World Bank..

An Open Digital Library Ordering System

Sarantos Kapidakis¹ and Kostas Zorbadelos¹

NHRF-National Hellenic Research Foundation,
NDC-National Documentation Center,
Vas. Constantinou Ave. PC 11635, Athens, Hellas (Greece).
Tel: +30-1-7273951,959. Fax: +30-1-7231699
{sarantos,kzorba}@ekt.gr

Abstract. In this paper we briefly describe an open ordering manipulation system with a WWW interface. The orders concern articles of scientific journals. Customer users can search in data sources from a variety of suppliers for articles of journals and order specific pages of the articles. Their search can also include electronic journals in which case their orders can be fulfilled, charged and delivered electronically as an e-mail attachment without needing an operator. The various suppliers can view orders made to them and service them. A customer can direct his order to several suppliers declaring an order of preference. We also introduce the issues involved and present our open system solution that separates the search from the order procedures. Searching can use any external interface provided by the various data sources and intercepts queries and their answers to search requests.

1 Introduction

The National Documentation Center (NDC) is a governmental infrastructure for providing information on research and technology. We provide a journal article ordering service, allowing certain individuals or organizations to search through a set of diverse external data sources containing information about journals and their contents and order them from the corresponding supplier(s). A typical usage scenario of the ordering system is described below.

A customer user (customer) searches a collection containing information about journals and their suppliers. The journal collection can also contain electronic journals. The search is done using the search interface provided by each data source. The answer to his request is a set of journals that match the search criteria and it also contains the corresponding suppliers.

After the correct journal is selected, the customer has the opportunity to place one or more linked orders for the specified item. The first order is tried first. If it concerns an electronic journal, it is serviced immediately, otherwise it is sent to the appropriate supplier. The supplier must manually respond to the customer's request. In the meanwhile, the customer can keep track of the status of his pending orders or state that he received the ordered item.

From a supplier's point of view the system looks as follows. A supplier is able to view the orders waiting to be served by him and either accepts or rejects

them. He also has the possibility to view any completed orders for statistics or reporting. To prevent a supplier from withholding and not responding to orders, we include a timeout mechanism that makes the next linked order active after a specified time interval.

2 Description of the Ordering System

There are various issues involved in the above-described procedures. First of all, a major issue is the diversity of the available data sources containing information about the articles, journals and suppliers, or the electronic journals. Most of these sources already have their own search interfaces and provide different functionality, something we exploit. Unification of these sources in a single format is sometimes possible but in other cases is a difficult, if not impossible task. Furthermore, since multiple suppliers, customers and orders are involved, authentication and security mechanisms become a necessity. On the other hand, orders should be stored in a central place for the extraction of various kinds of statistical and financial reports. Last but not least, the system should have an easy-to-use intuitive interface as it is designated for use by librarians and other people not having any technical background in computers. In the following sections we describe the ordering system's functionality which addresses those issues and is also extensible, using many different external data sources.

The Figure 1 shows in general the system's functionality.

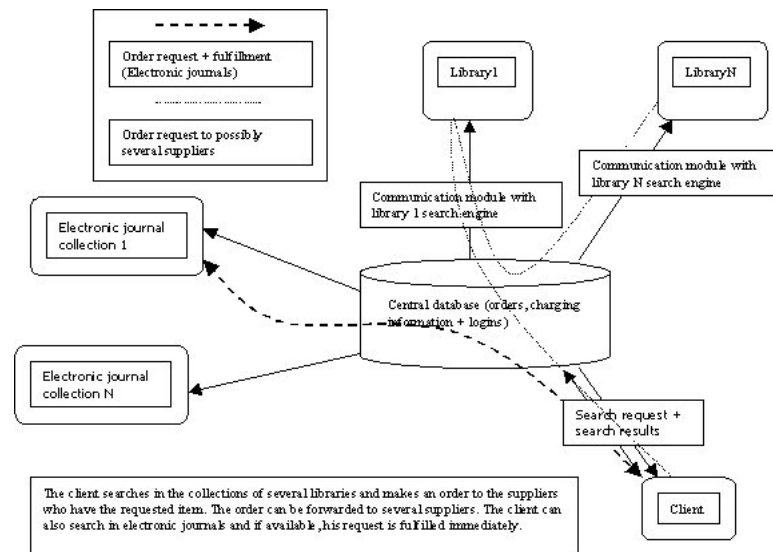


Fig. 1. General system architecture

2.1 Customer Search and Selection

The first thing a customer must do is search in the available data sources. In an open system, we do not store any information regarding the target material, journals or articles, in the central database. All such information is accessed from external sources and we use them by their search interfaces to obtain the results concerning a user's request. The system only needs to know the communication protocol with each specified external system. We can intercept calls to Z39.50 servers which give us access to collections of various libraries with the Z39.50 protocol, or to the HTTP protocol of the web interfaces and search engines of various electronic journals. We use proxying techniques to intercept the various queries and answers from the different systems and present them to the user. Thus, data are updated easily by their owners (suppliers) on their servers. We use read-only access.

The customer is prompted with the set of external sources available and is given the chance to choose one for his search. After that, he is presented the search screen of the selected source with various searchable fields. The customer can search for articles wherever available and find the journals that contain them as well as the suppliers that have them. In each case, the answer contains the specific item and the corresponding suppliers. The customer can then initiate the order procedure.

2.2 Ordering

After the search and selection of the requested journal the customer can create a linked order list to several possible suppliers. This step requires login and password authentication. In the case where the search included article information all the available data for the order are gathered, whereas in the case where article information is not available the customer is prompted to enter any missing fields, such as issue, article, start and end pages etc. The first order in the list is marked active and an agent hook executes. The agent hook executes each time an order in the list becomes current and the order is passed as a parameter. The actions of the hook depend on the type of the corresponding supplier. This way all order manipulation in the system is uniform.

Robot Supplier. In the case of electronic journals the completion of the order is made immediately and the requested pages are sent electronically as an e-mail attachment. In this case, the hook is responsible for fulfilling the order, that is locate, charge and electronically send the ordered article. It also marks the status of the order completed. We must note that in the case of electronic subscriptions the server computer is the only computer that needs access to the electronic journal data and handles and records their usage. Access to such data is often IP restricted and we record any usage of them according to the contract with each publisher. Other computers cannot directly search and retrieve any electronic journal data if any restriction is imposed by the publisher.

Manual Supplier. In the case of a manual supplier the hook just makes the corresponding order in the list active and a response from the supplier is expected manually. If the supplier rejects the request (e.g. in the case he can't serve the request) or does not answer for a specified period of time (timeout) the order is automatically forwarded to the next supplier making the next order in the list active and triggering the agent hook all over again.

The manual supplier is first logged into the system and browses the orders he must serve. Supplier logins can be IP restricted as an extra security feature. He is presented a list of pending orders and after viewing the details of each one he is able to reply. First of all, he declares whether he accepts or rejects the order and for each accepted order he also fills or corrects various fields such as number of pages, delivery method etc.

2.3 Statistics

Since all data regarding orders are stored in a central database, various statistics can easily be extracted. We can have per customer statistics where each customer can see total cost of completed orders, mean response time of the various suppliers or any other interesting data. We can also have per supplier reports that show for example the entire history of serviced orders. Several financial reports or valuable statistics regarding system usage can also be extracted for all users of the system.

3 Advantages - Future Extensions

Our system makes distributed operation and update possible because it avoids the unification of data in one form, something that in many cases is impossible, and it exploits the various search engines of the outside systems. No single entity is responsible for the gathering and update of the journal data but each data source is responsible for the data it provides. Moreover the system can easily be extended to support other data sources only by adding support for the protocol used to communicate with the source. The central database holding the orders themselves remains intact. We also use different algorithms for cost evaluation of orders. The charging of orders is a different customizable module in the system. Finally each customer can view the status of his pending orders.

As an alternative to frequent users, with login, we could have one-time users, where no logins are required if we provide for credit card charging. Of course this means that secure protocols must be employed for the exchange of such sensitive data. We could also provide certain private collections and allow each customer user to choose among them to restrict his search. For instance, we could have a collection of computer science journals and another of medicine journals.

Special NKOS Workshop on Networked Knowledge Organization Systems

Martin Doerr¹, Traugott Koch², Douglas Tudhope³, and Repke de Vries⁴

¹ Institute of Computer Science, Foundation for Research and Technology,
Hellas (FORTH), Heraklion, Greece.
martin@csi.forth.gr

² NetLab, Lund University Library Development Department, Sweden.
traugott@munin.lub.lu.se

³ School of Computing, University of Glamorgan, Pontypridd, Wales, UK.
dstudhope@glamorgan.ac.uk

⁴ Netherlands Institute for Scientific Information Services, Netherlands.
repke.de.vries@niwi.knaw.nl

Objectives

This half-day workshop aims to provide an overview of research, development and projects related to the usage of knowledge organization systems in Internet based services and digital libraries. These systems can comprise thesauri and other controlled lists of keywords, ontologies, classification systems, taxonomies, clustering approaches, dictionaries, lexical databases, concept maps/spaces, semantic road maps etc.

A second objective of the workshop is to enable and support co-operation between European initiatives in the area of networked knowledge organization and to provide a basis for participation in global efforts (see below for other NKOS projects) and standardization processes.

This workshop represents a chance to reach out to a broader group of people working in the area, to inform each other on on-going research and projects and to start discussions about possible common goals and tasks, including organizational efforts: e.g. setting up a regular event and communication or collaboration with global NKOS activities.

Workshop Content and Structure

The half-day workshop spans two conference sessions on September 20. The workshop will start with an introduction and short statements of experience and interests from all participants. The main content of the first session will be presentations from 4 invited panel speakers:

OIL: The Ontology Inference Layer

Sean Bechhofer, Computer Science Department, University of Manchester, UK.
<http://potato.cs.man.ac.uk/seanb/>

Thesaurus Mapping

Martin Doerr, Foundation for Research and Technology, Hellas (FORTH), Greece.
<http://www.ics.forth.gr/>

Report on NKOS activities in North America

Linda Hill, Alexandria Digital Library Project, University of California, Santa Barbara, USA.
<http://www.alexandria.ucsb.edu/~lhill/nkos>

Having the Right Connections: The Limber Project

Ken Miller, Data Archive, University of Essex, UK
<http://www.cordis.lu/ist/projects/99-11748.htm>

After lunch, the second workshop session will offer an opportunity for all participants to informally contribute position statements, research topics, project experiences and perspectives, leading into general (or optionally small-group) discussion. The workshop will conclude by considering options for cooperation and future activities.

To facilitate participation in the workshop, we encourage all interested participants to email short position statements to any of the workshop organisers. These will be posted on ECDL2000 section of the NKOS website:
<http://www.alexandria.ucsb.edu/~lhill/nkos/>

Overview

In recent years, the need for knowledge organization in Internet services supporting resource discovery in digital libraries and related areas has been increasingly recognised. The growth of information on the Web continues to challenge Internet searchers to locate relevant, high quality resources. The major Web search services respond to this challenge by increasing the size of their databases and offering more powerful searching and ranking features. In contrast to these global search engines, many smaller, more discriminating services are trying to improve resource discovery on the Internet by focusing their efforts on the selection, description, and subject arrangement of high-quality resources.

Most quality-based services go beyond selection and also catalog or describe chosen resources according to metadata standards such as the Dublin Core (DC). Quality-based services also commonly provide subject access to selected resources through formal knowledge organization structures such as a subject thesaurus, classification scheme, or both. Beyond the insufficiencies of existing individual services, especially the big search services, the overall resource discovery infrastructure on the Internet is not very well developed. For example, few support tools exist for assisting users in finding the right service to start with. Connections among subject collections are also lacking. A user who discovers one service, e.g., about mechanical engineering, is not normally forwarded to related topics in another service. There is no discovery architecture or linkage among services based on subject indexing languages or domain ontologies.

Progress in the distributed usage of networked knowledge organization systems may well provide the key to good solutions for these resource discovery problems and be instrumental in dealing with the complexity of subject access to distributed digital information.

The issue poses questions of pure research, but also to a very high degree applied questions of agreement, collaboration, standardization, use and user needs. Problems of pure research concentrate around ontological questions of structuring and connecting large, multipurpose knowledge bases and on questions of advanced representation and reasoning mechanisms. The applied problems are characterized by the lack of data exchange format standards, communication protocols and reference architectures, in general a result of insufficient awareness of the relevance of global formal knowledge resources to access distributed, diverse and heterogeneous sources.

Workshop Topics

Digital Library Requirements for Knowledge Organization Schemas:

- The need for knowledge organization in subject gateways and discovery services, issues of application and use
- Web-based directory structures as knowledge organization systems
- Knowledge organization as support for web-based information retrieval, query expansion, cross-language searching
- Semantic portals

Digital Library Requirements for Knowledge Based Data Processing:

- Knowledge organization for filtering, information extraction, summary
- Knowledge organization support for multilingual systems, natural language processing or machine translation
- Structured result display, clustering
- End-user interactions with knowledge organization systems, evaluation and studies of use, knowledge bases for supportive user interfaces, visualization

Digital Library Requirements for Knowledge Structuring and Management:

- Suitable vocabulary structures, conceptual relationships
- Comparison between established library classification systems and home-grown browsing structures
- Methodologies, tools and formats for the construction and maintenance of vocabularies and for mapping between terms, classes and systems
- Frameworks for the analysis of assumptions and viewpoints underlying the construction and application of terminology systems
- Methods for the combination and adaptation of different vocabularies

Digital Library Requirements for Access to Knowledge Structures:

- Data exchange and description formats for knowledge organization systems, the potential and limitations of XML and RDF schemas
- Handling of subject information in metadata formats
- Standards and repositories for machine-readable description of networked knowledge organization schemas (as collections/systems)
- Interoperability, cross-browsing and cross-searching between distributed services based on knowledge organization systems

- Distributed access to knowledge organization systems: standard solutions and protocols for query and response, taxonomy servers

Communities Involved:

NKOS: Networked Knowledge Organization Systems

<http://www.alexandria.ucsb.edu/~lhill/nkos/>

Projects include content standards for describing networked knowledge organization systems and development of a model for a protocol for NKOS query and response. The aim is to develop cooperation between researchers and developers who are creating networked interactive knowledge organization systems. NKOS has held three previous workshops in the USA at ACM DL conferences – this is the first European workshop on the topic.

- IMesh: International Collaboration on Internet subject gateways
<http://www.imesh.org/>
- The NORDIC Metadata Projects
<http://linnea.helsinki.fi/meta/>
- MODELS (MOving to Distributed Environments for Library Services) Terminology Workshop
<http://www.ukoln.ac.uk/dlis/models/>
- The topic is also relevant to digital library, museum, archives, and cultural heritage communities, geo-spatial research, systems for geo-referencing. Related topics have been/are addressed (for example) in:
 - the EU projects for Telematics and their successors:
Aquarelle, Term-IT, DESIRE, Renardus;
LIMBER, and others;
 - various National Digital Library initiatives:
RDN, DEF Denmark, Finland, Netherlands etc.
 - various Standardization initiatives:
Electronic thesauri (NISO workshop)
Zthes Z39.50; Dublin Core Metadata Initiative; RDF Schema

Expected Participants

- Digital library and information infrastructure developers
- Resource discovery service providers (search engines, directories, subject gateways, portals)
- Information scientists, library, museum and archive professionals
- Thesaurus and ontology developers
- Standard developers in the area of terminology usage and exchange
- Computer scientists, tool developers, interface designers
- Knowledge managers

Implementing Electronic Journals in the Library and Making them Available to the End-User: An Integrated Approach

Gerrit Alewaeters¹, Serge Gilen¹, Paul Nieuwenhuysen¹, Stefaan Renard¹, and Marc Verpoorten¹

¹ Vrije Universiteit Brussel, University Library, Pleinlaan 2,
1050 Brussel, Belgium
{galewaet, sgilen, pnieuwen, strenard, mverpoort}@vub.ac.be

Abstract. This short-paper describes our strategy for implementing electronic journals (with embedded multimedia) in the library and making them available to the end-user. Together with the Technische Universiteit Eindhoven (TUE) in the Netherlands, we own the source code of the Vubis library information system, which allows development, customization and tighter integration of Vubis for our specific needs. The university library of the Vrije Universiteit Brussel (VUB) in Belgium is currently involved in a digital library project named CROCODIL (CROss-platform CO-operation for a DIgital Library 1999-2000) which is sponsored by the Flemish government (IWT). Our partners are the largest subscription agent in Europe (Swets Blackwell) and the distributor of the Vubis library information system (Geac). We are developing integrated access to electronic documents (with embedded multimedia) and information in different formats. We hope that our experience can be of interest to other libraries coping with the integration of electronic journals in their library system.

1 The Context: Strategies for Implementing Electronic Journals

One of the challenges libraries are facing is providing adequate access to heterogeneous information sources: traditional paper collections and digital collections. The so-called hybrid library tries to bring both worlds (print and digital) tighter together [1]. More specifically, we present a method to integrate electronic journals in the library information system.

The opac (online public access catalogue) has been for years the source to search information on print journals (holdings and location). With the breakthrough of the Web, electronic journals are booming. CrossRef has been launched in February 2000; it has the potential to be a killer application for electronic journals, enabling links from references in a full-text article utilizing digital object identifiers (DOI) [2]. All this poses new challenges to the library: should we integrate electronic journals in the existing library information system or should we follow separate routes for them? Several strategies exist today for implementing electronic journals:

Some libraries opt for solutions which are not well integrated in the existing library information system:

- web-based static listing by title, publisher and/or subject
- local searchable database of titles
- gateway of subscription agents and publishers who act as aggregator
- full-text database

Other libraries integrate electronic journals in their library information system more tightly using one-level linking or multi-level linking. One-level linking provides a basic link to the electronic journal (title level or publisher level); listings by title, publisher and/or subject can be generated from this system. Multi-level linking is our strategy, it will be explained in detail.

2 A Strategy to Implement Electronic Journals

The Vubis catalogue named Article Database (ADB) is a local implementation of SwetScan, that contains the table of contents of the most important international scholarly journals. The database is updated daily via Swets Blackwell by ftp on our computer, then the SwetScan data is converted to the Vubis format and indexed (journal: title, title word, ISSN, and article: title word(s), author name (and/or first name initial), publication date). Holding information is presented in ADB for the traditional print journals to which the library subscribes. At the article level a link to the interlibrary lending (ILL) module of the Vubis library system is shown. When a library user clicks the ILL link, the system asks for authentication by library card barcode and password. After authentication a form, generated from the bibliographical record of the requested article in ADB, is forwarded from the ILL module to the interlibrary lending department, where the description can serve as part of the input in Impala. Impala provides Belgian libraries with a centralized system for sending and receiving (ILL) requests [3]. The integration of full-text links in ADB was our next goal.

The library subscribes to a few thousand journals which have free or paid electronic counterparts. The library had already access to several electronic journals from different publishers, but not all through SwetsNet [4]. In addition SwetsNet was a stand-alone service in the library, there was no integration with the Vubis library information system. The library policy is to provide access to only those electronic journals which use authentication by recognition of IP range. Unfortunately some key journals, like *The Lancet*, are working with authentication by username and password, but this is not suitable for a public access system. Our partner Swets Blackwell does not know how to solve this problem.

Electronic titles included in SwetsNet

Besides access through Swetsnet, Swets Blackwell also allows integration of an opac with electronic journals through the so-called multi-level linking [5]. This works through a link manager of Swets Blackwell that has a simple algorithm to build a URL, based only upon the information found in a typical bibliographical record:

< <a href="http://www.swetsnet.nl/link/access_db?issn=<...>&vol=<...>&iss=<...>page=<...>">http://www.swetsnet.nl/link/access_db?issn=<...>&vol=<...>&iss=<...>page=<...> >	
-----	-----
fixed section	derived section

Only the first part of the URL is fixed; the second part after the question mark can be derived from bibliographical information (derived section). With multi-level linking we can link directly from the article description in ADB to the abstract level in SwetsNet. The SwetsNet links are delivered with the SwetScan data and subscribed titles are activated in ADB. So far so good, but this new service in ADB missed holdings information of the electronic journals. For example, an electronic journal is only available from 1999 until March 2000. Even worse: there can be gaps in the availability of electronic journals; e.g. issues from 1997 onwards are available but not from 1998. The system needs these ranges if we only want to present links which are really leading to full-text. One solution could be to let library staff fill in the full-text ranges in the Vubis library information system gathered from SwetsNet webpages. However there are several drawbacks. First of all it is labour intensive and involves a human effort which is not free of errors. Secondly these ranges are not static; library staff should check them regularly. An improvement may be to harvest automatically these ranges regularly from the SwetsNet webpages.

Electronic titles not included in SwetsNet

Non SwetsNet titles were gathered from information Swets Blackwell delivered to us in an Excel file. Every potential electronic journal had to be checked manually on authentication method by library staff. As many electronic journals as possible are activated and incorporated in ADB. Some publishers have link managers like Swets Blackwell: e.g. American Institute of Physics, American Physical Society, CERN SetLink. Link managers are library friendly, because the burden of maintaining the URLs is completely in the hands of the content provider. Since we are still dealing with URLs, reliability is a major issue. Hopefully more content providers will launch link managers for their electronic journals collection. URLs of non SwetsNet titles are deciphered by publisher. The derived section of the URL is generated dynamically from ADB using the M computer language (MUMPS). Holdings information can be harvested from the electronic journals sites.

3 Viewing Electronic Journals

Most full-text articles are delivered in HTML or in Portable Document Format (PDF). A growing number of electronic journals include multimedia information. In the area of natural sciences the potential for the use of multimedia is high [6].

To keep users and certainly managers of public Internet workstations up to date there is a need for some overview on a Website of important file formats with test files and corresponding display software. Several sites about file formats have been tested by us, but no one was up to our expectations [7]. Therefore, we plan to create such a tool.

References

1. Brophy, P.: Towards a generic model of information and library services in the information age. *Journal of Documentation* 2 (2000) 161-184
2. CrossRef FAQ [Online]. [cited 2000-04-24]. Available from Internet: <<http://www.crossref.org/faqs.htm>>.
3. Impala. The Belgian electronic document ordering system [Online]. [cited 2000-04-18]. Available from Internet: <http://lib.ua.ac.be/IMPALA/DOCSTORE/folder_e.html>
4. <<http://www.swetsnet.nl>>
5. Multi-level Linking in SwetsNet [Online]. [cited 2000-04-18]. Available from Internet: <<http://www.nesli.ac.uk/linking.html>>.
6. Eason, K. (et. al.): A Comparative Analysis of the Role of Multi-Media Electronic Journals in Scholarly Disciplines. Final Report, November 1997, eLib Supporting Study. [Online]. [cited 2000-04-24]. Available from Internet: <http://www.ukoln.ac.uk/services/elib/papers/supporting/pdf/eason.pdf>
7. Plug-in Gallery & Demo Link <<http://www2.gol.com/users/oyamada/>>, BrowserWatch Plug-In Plaza <<http://browserwatch.internet.com/plug-in.html>>, Netscape Plug-in <<http://home.netscape.com/plugins/>>, Multimedia File Formats on the Internet: A beginner's Guide for PC Users <<http://www.lib.rochester.edu/multimed/contents.htm>>, What Is...every file format in the world <<http://www.whatis.com/ff.htm>>.

Author Index

- Adachi, Jun 35
Afifi, Marianne 481
Aguilera Caballero, Begoña 385
Al-Hawamdeh, Suliman 14
Alani, Harith 45
Alewaeters, Gerrit 506
Amaral, A. Maia do 1
Amaral, Rui 315
Amato, Giuseppe 328
Anan, Hesham 168
Androutsopoulos, Ion 403
Apps, Ann 93
Atarashi, Rei S. 456
- Baccarelli, Ricardo 449
Bainbridge, David 249
Barbosa, Eduardo 449
Barbosa, Maurício 449
Barrueco Cruz, José Manuel 127
Bauer, Michael G. 180
Becker, Hans 424
Belisle, Claire 239
Berrut, Catherine 206
Bicudo, Rose 449
Bielec, John A. 413
Böhm, Klemens 83
Bolanho, Andréa 449
Bollen, Johan 356
Brickley, Dan 103
Brunning, Dennis 475
Buchanan, George 249
- Caldas Pinto, J.R. 1, 24
Campos, João P. 380
Cassidy, Noel 249
Castelli, Donatella 328
Chandrinos, Konstantinos V. 403
Chihara, Kunihiro 456
Cooper, Brian 128
Costa, António 376
Costa, Dilson da 449
Covert, Kay 399
Crespo, Arturo 128, 190
Crestani, Fabio 324
Cunningham, Sally Jo 418
- Dalhousmi, Salah 239
Decker, Stefan 490
Derniame, Jean-Claude 494
Díaz Esteban, Alberto 259
Dittenbach, Michael 348
Doerr, Martin 502
Drobnik, Oswald 117
Drori, Offer 69
Dzeakou, Patricia 494
- Esposito, F. 229
- Fanizzi, N. 229
Ferilli, S. 229
Freitas, Vasco 376
Fuhr, Norbert 352, 364
Fujino, Asa 449
- Galavotti, Luigi 59
Gameiro, Fátima 449
García Jiménez, Antonio 259
Garcia-Molina, Hector 128, 190
Gergatsoulis, Manolis 368
Gibb, F. 305
Gilen, Serge 506
Gövert, Norbert 352
Gómez-Navarro, Pablo Gervás 259
Gonzales, Camila 449
Granado, I. 1
Großjohann, Kai 364
Guimarães, José 449
- Hansen Montgomery, Carol 413
Harada, Elena 449
Hembise, Céline 445
Herrington, Scott 471
Heß, Martin 117
Hill, Linda L. 280
Hoppenbrouwers, Jeroen 467
Huang, Michael B. 441
Hunter, Jane 103, 291
Huxley, Lesly 395
- Imai, Masakazu 456

- James, Darren 291
Jannink, Jan 490
Jones, Christopher 45
Jones, Gareth J. F. 340
Jones, Matt 249
- Kapidakis, Sarantos 498
Katata, Tadashi 456
Khalil, Mounir A. 360
Khan, Gul N. 14
Klas, Claus-Peter 352
Klink, Markus J. R. 127
Kobayashi, Sérgio 449
Kobashi, Nair 449
Koch, Traugott 502
Koehler, Wallace C. 407
Kokkelink, Stefan 364
Konomos, Philip 471
Kovács, László 372
Krichel, Thomas 127
- Lagoze, Carl 103, 144
Landoni, Monica 305, 485
Lara, Marilda 449
Larouk, Omar 239
Lee, Hyowon 206
Lehner, Richard 385
Leme, Paulo 449
Li, Guohui 441
Lim, Tek Yong 389
- Macedo, Joaquim 376
MacIntyre, Ross 93
Mahoui, Malika 418
Maly, Kurt 168
Marcolino, A. 1
Maria, Nuno 344
Marlow, Séan 206
Matthes, Florian 269
Melnik, Sergey 490
Mengucci, M. 1, 24
Merkel, Dieter 348
Mezzeti, Dilon 449
Micsik, András 372
Mitakos, Theodoros 368
Mitra, Prasenjit 490
Mönch, Christian 117, 158
Muge, F. 1, 24, 319
Mukaiyama, Hiroshi 336
Murphy, Noel 206
- Naime, Ana 449
Nasir, Norliza Mohd 249
Nichols, David M. 239
Niederée, Claudia 269
Nieuwenhuysen, Paul 506
Nunes, Márcio 449
- O'Connor, Noel E. 206
Ohta, Manabu 35
- Paliouras, Georgios 403
Pasquier, Florent 219
Pataki, Balázs 372
Payette, Sandra 144
Pelizzari, Eugenio 452
Pemberton, Duncan 239
Pereira, Fernando 332
Pereira Filho, Innocência 449
Peterson, Elaine 437
Pina, P. 1, 24
Pisani, Serena 328
Place, Thomas 467
Plonski, Guilherme 449
Prado, Cláudia 449
- Quested, David J. 340
- Ramalho, Mário 1, 24
Ramos, Victorino 1, 319
Rauber, Andreas 348
Reis, Ana 449
Renard, Stefaan 506
Rocha, Luis M. 356
- Schek, Hans-J. 83
Schmidt, Joachim W. 269
Sebastiani, Fabrizio 59
Semeraro, G. 229
Shiratuddin, Norshuhada 485
Silva, Mário J. 344, 380
Simi, Maria 59
Sirakov, N. 1
Smeaton, Alan F. 206
Specht, Günther 180
Spyropoulos, Constantine D. 403
Staab, Steffen 490
Stavrakas, Yannis 368
Steffens, Ulrike 269
Studer, Rudi 490
Sunahara, Hideki 456

Takasu, Atsuhiko	35	Weber, Roger	83
Tan, Dun	168	Wegner, Bernd	424, 428, 461
Tang, Enya Kong	389	Wiederhold, Gio	490
Theng, Yin Leng	249	Wilson, R.	305
Thimbleby, Harold	249		
Thomson, Katherine E.	340	Yahia, Bakelli	432
Trancoso, Isabel	315	Yoshizawa, Erina	449
Tudhope, Douglas	45, 502		
Twidale, Michael B.	239	Zhang, Yunchuan	168
		Zibreira, Carla	332
Verpoorten, Marc	506	Zorbadelos, Kostas	498
Vieira, P.	1, 24	Zsámboki, István	372
Vogel, Michely	449	Zubair, Mohammad	168
Vries, Repke de	502		