Julio Gonzalo
Costantino Thanos
M. Felisa Verdejo
Rafael C. Carrasco (Eds.)

LNCS 4172

Research and Advanced Technology for Digital Libraries

10th European Conference, ECDL 2006
Alicante, Spain, September 2006
Proceedings

Springer

Lecture Notes in Computer Science 4172

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Julio Gonzalo Costantino Thanos
M. Felisa Verdejo Rafael C. Carrasco (Eds.)

Research and Advanced Technology for Digital Libraries

10th European Conference, ECDL 2006
Alicante, Spain, September 17-22, 2006
Proceedings

Springer

Volume Editors

Julio Gonzalo
M. Felisa Verdejo
Universidad Nacional de Educación a Distancia (UNED)
Departamento de Lenguajes y Sistemas Informáticos
c/Juan del Rosal, 16, 28040 Madrid, Spain
E-mail:{julio,felisa}@lsi.uned.es

Costantino Thanos
Consiglio Nazionale delle Richerche
Istituto di Scienza e Tecnologie dell'Informazione
Via Moruzzi, 1, 56124, Pisa, Italy
E-mail: Costantino.Thanos@isti.cnr.it

Rafael C. Carrasco
Universidad de Alicante
Departamento de Lenguajes y Sistemas Informáticos
03071 Alicante, Spain
E-mail: carrasco@dlsi.ua.es

Library of Congress Control Number: Applied for

CR Subject Classification (1998): H.3.7, H.2, H.3, H.4.3, H.5, J.7, J.1, I.7

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

ISSN	0302-9743
ISBN-10	3-540-44636-2 Springer Berlin Heidelberg New York
ISBN-13	978-3-540-44636-1 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springer.com

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11863878 06/3142 5 4 3 2 1 0

Preface

We are proud to present the proceedings of the 10th European Conference on Digital Libraries (ECDL 2006) which, following Pisa (1997), Heraklion (1998), Paris (1999), Lisbon (2000), Darmstadt (2001), Rome (2002), Trondheim (2003), Bath (2004) and Vienna (2005), took place on September 17-22, 2006 at the University of Alicante, Spain. Over the last ten years, ECDL has created a strong interdisciplinary community of researchers and practitioners in the field of digital libraries, and has formed a substantial body of scholarly publications contained in the conference proceedings. As a commemoration of its 10th anniversary, and by special arrangement with Springer, these proceedings include (as an attached CD-ROM) an electronic copy of all ECDL proceedings since its inception in 1997: a small but rich digital library on digital libraries.

ECDL 2006 featured separate calls for paper and poster submissions, resulting in 130 full papers and 29 posters being submitted to the conference. All papers were subject to an in-depth peer-review process; three reviews per submission were produced by a Program Committee of 92 members and 42 additional reviewers from 30 countries. Finally, 36 full paper submissions were accepted at the Program Committee meeting, resulting in an acceptance rate of 28%. Also, 15 poster submissions plus 18 full paper submissions were accepted as poster or demo presentations, which are also included in this volume as four-page extended abstracts.

The conference program started on Sunday 17 with a rich tutorials program, which included a tutorial on thesauri and ontologies in digital libraries by Dagobert Soergel, and introduction to digital libraries by Ed Fox, a tutorial on bringing digital libraries to distributed infrastructures by Thomas Risse and Claudia Niederée, a description of the Fedora repository and service framework by Sandy Payette and Carl Lagoze, a tutorial on creating full-featured institutional repositories combining DSpace ETD-db and DigiTool, and a tutorial on the use of XML and TEI for content production and metadata.

The main conference featured three keynote speakers: Michael A. Keller, Ida M. Green University Librarian at Stanford, Director of Academic Information Resources, Publisher of HighWire Press, and Publisher of the Stanford University Press; Horst Forster, director of *Interfaces, knowledge and content technologies* at the Directorate-General for Information Society of the European Commission, and Ricardo Baeza-Yates, director of Yahoo! Research Barcelona and Yahoo! Research Latin America at Santiago de Chile.

The rest of the main conference program consisted of 12 technical sessions, a panel and a poster session preceded by a spotlight session which served as a quick guide to the poster session for the conference participants.

Following the main conference, ECDL hosted eight workshops, including the long-standing workshop of the Cross-Language Evaluation Forum, a major event of its own that ran an intensive three-day program devoted to discuss the outcome of its annual competitive evaluation campaign in the field of Multilingual Information Access. The other workshops were: NKOS 2006 (5th European Networked Knowledge Organization Systems workshop), DORSDL 2006 (Digital Object Repository Systems in Digital Libraries), DLSci 2006 (Digital Library goes e-science: perspectives and challenges), IWAW 2006 (6^{th} International Workshop on Web Archiving and Digital Preservation), LODL 2006 (Learning Object repositories as Digital Libraries: current challenges), M-CAST 2006 and CSFIC 2006 (Embedded e-Learning – critical success factors for institutional change). All information about ECDL 2006 is available from the conference homepage at http://www.ecdl2006.org.

We would like to take the opportunity to thank all those institutions and individuals who made this conference possible, starting with the conference participants and presenters, who provided a dense one-week program of high technical quality. We are also indebted to the Program Committee members, who made an outstanding reviewing job under tight time constraints; and to all Chairs and members of the Organization Committee, including the organizing teams at the University of Alicante, Biblioteca Virtual Miguel de Cervantes and UNED. We would specifically like to thank Miguel Ángel Varó for his assistance with the conference management system, and Valentín Sama for his help when compiling these proceedings.

Finally, we would also like to thank the conference sponsors and cooperating agencies: the DELOS network of Excellence on Digital Libraries, Grupo Santander, Ministerio de Educación y Ciencia, Patronato Municipal de Turismo de Alicante, Red de investigación en Bibliotecas Digitales, Fundación Biblioteca Miguel de Cervantes, Departamento de Lenguajes y Sistemas Informáticos de la Universidad de Alicante, and UNED.

Julio Gonzalo
Costantino Thanos
Felisa Verdejo
Rafael Carrasco

Organization

Organization Committee

General Chair

Felisa Verdejo	UNED, Spain

Program Chairs

Julio Gonzalo	UNED, Spain
Costantino Thanos	CNR, Italy

Organization Chair

Rafael C. Carrasco	Universidad de Alicante, Spain

Workshops Chairs

Donatella Castelli	CNR, Italy
José Luis Vicedo	Universidad de Alicante, Spain

Poster and Demo Chair

Günter Mühlberger	University of Innsbruck, Austria

Panel Chairs

Andreas Rauber	Vienna University of Technology, Austria
Liz Lyon	UKOLN, UK

Tutorial Chairs

Marcos Andre Gonçalves	Federal University of Minas Gerais, Brazil
Ingeborg Solvberg	Norwegian University of Science and Technology, Norway

Publicity and Exhibits Chairs

Maristella Agosti	University of Padua, Italy
Tamara Sumner	University of Colorado at Boulder, USA
Shigeo Sugimoto	University of Tsukuba, Japan

Doctoral Consortium Chairs

Jose Borbinha	IST, Lisbon Technical University, Portugal
Lillian Cassel	Villanova University, USA

Local Organization Chair

Rafael C. Carrasco	University of Alicante

Program Committee

Alan Smeaton	Dublin City University, Ireland
Allan Hanbury	Vienna University of Technology, Austria
Andras Micsik	SZTAKI, Hungary
Andy Powell	Eduserv Foundation, UK
Anita S. Coleman	University of Arizona, USA
Ann Blandford	University College London, UK
Anselmo Peñas	UNED, Spain
Antonio Polo	University of Extremadura, Spain
Birte Christensen-Dalsgaard	State and University Library, Denmark
Boris Dobrov	Moscow State University, Russia
Carl Lagoze	Cornell University, USA
Carlo Meghini	ISTI-CNR, Italy
Carol Peters	ISTI-CNR, Italy
Ching-Chih Chen	Simmons College, USA
Christine L. Borgman	University of California, USA
Clifford Lynch	Coalition for Networked Information, USA
Dagobert Soergel	University of Maryland, USA
Dieter Fellner	Graz University of Technology, Austria
Dimitris Plexousakis	FORTH, Greece
Djoerd Hiemstra	Twente University, Netherlands
Donna Harman	NIST, USA
Douglas W. Oard	University of Maryland, USA
Eduard A. Fox	Virginia Tech, USA
Edleno Silva de Moura	Universidade do Amazonas, Brazil
Ee-Peng Lim	Nanyang Technological University, Singapore
Elaine Toms	Dalhousie University, Canada
Erik Duval	Katholieke Universiteit Leuven, Belgium
Fernando López-Ostenero	UNED, Spain
Floriana Esposito	University of Bari, Italy
Franciska de Jong	University of Twente, Netherlands
Frans Van Assche	European Schoolnet, Belgium
Gary Marchionini	University of North Carolina Chapel Hill, USA
George Buchanan	University of Wales, Swansea, UK

Gerhard Budin	University of Vienna, Austria
Gregory Crane	Tufts University, USA
George R. Thoma	U.S. National Library of Medicine, USA
Hanne Albrechtsen	Institute of Knowledge Sharing, Denmark
Harald Krottmaier	Graz University of Technology, Austria
Heiko Schuldt	University of Basel, Switzerland
Herbert Van de Sompel	Los Alamos National Laboratory, USA
Howard Wactlar	Carnegie Mellon University, USA
Hussein Suleman	University of Cape Town, South Africa
Ian Witten	University of Waikato, New Zealand
Ingeborg Solvberg	Norwegian University of Technology and Science, Norway
Jacques Ducloy	CNRS-INIST, France
Jan Engelen	Katholieke Universiteit Leuven, Belgium
Jane Hunter	University of Queensland, Australia
Jela Steinerova	Comenius University in Bratislava, Slovakia
Jesús Tramullas	University of Zaragoza, Spain
José Hilario Canós Cerdá	Universidad Politécnica de Valencia, Spain
Jussi Karlgren	SICS, Sweden
Key-Sun Choi	Korea Advanced Institute of Science and Technology, Korea
Laurent Romary	Laboratoire Loria CNRS, France
Lee-Feng Chien	Academia Sinica, Taiwan
Leonid Kalinichenko	Russian Academy of Sciences, Russia
Liddy Nevile	La Trobe University, Australia
Lloyd Rutledge	CWI, Netherlands
Lynda Hardman	CWI, Netherlands
Marc Nanard	University of Montpellier, France
Margaret Hedstrom	University of Michigan, USA
Margherita Antona	FORTH, Greece
Mária Bieliková	Slovak University of Technology in Bratislava,, Slovakia
Maria Sliwinska	ICIMSS, Poland
Mario J. Silva	Universidade de Lisboa, Portugal
Martin Kersten	CWI, Netherlands
Michael Mabe	Elsevier, UK
Mike Freeston	University of California, Santa Barbara, USA
Mounia Lalmas	Queen Mary University of London, UK
Nicholas Belkin	Rutgers University, USA
Nicolas Spyratos	Université de Paris-Sud, France
Norbert Fuhr	University of Duisburg-Essen, Germany
Nozha Boujemaa	INRIA, France
Pablo de la Fuente	University of Valladolid, Spain
Paul Clough	University of Sheffield, UK
Rachel Bruce	JISC, UK
Ray R. Larson	University of California, Berkeley, USA

Reagan Moore	SDSC, USA
Ricardo Baeza-Yates	Yahoo! Research, Spain and Chile
Richard Furuta	Texas A&M University, USA
Sally Jo cunningham	University of Waikato, New Zealand
Sarantos Kapidakis	Ionian University, Greece
Schubert Foo	Nanyang Technological University, Singapore
Stavros Christodoulakis	Technical University of Crete, Greece
Stefan Gradmann	University Hamburg Computing Center, Germany
Susanne Dobratz	Humboldt University, Germany
Thomas Baker	State and University Library, Germany
Thomas Risse	Fraunhofer IPSI, Germany
Timos Sellis	National Technical University of Athens, Greece
Tiziana Catarci	University of Rome 1, Italy
Traugott Koch	UKOLN, UK

Additional Reviewers

Enrique Amigó
Alia Amin
Bruno Araújo
Luis J. Arévalo Rosado
Javier Artiles
David Bainbridge
Tobias Blanke
André Carvalho
Michelangelo Ceci
You-Jin Chang
Theodore Dalamagas
Reza Eslami
Nicola Fanizzi
Stefano Ferilli
Gudrun Fischer
Daniel Gomes
Sheila Gomes
Mark Hall
Jesús Herrera
Michiel Hildebrand
Stephen Kimani
Sascha Kriewel
Monica Landoni
Francesca A. Lisi
Manuel Llavador
Natalia Loukachevitch
Ming Luo
Jorge Martínez Gil
Roche Mathieu
Diego Milano
Lehti Patrick
Víctor Peinado
Thomaz Philippe
Antonella Poggi
Konstantin Pussep
Philippe Rigaux
Dimitris Sacharidis
Monica Scannapieco
Yannis Stavrakas
Maria Sliwinska
Zoltan Szlavik
Theodora Tsikrika

Local Organization Committee

Laura Devesa	(Office Contact)
Antonio Carrasco	(Coordination)
Rafael González	(Communication and Press)
Ester Serna	(Website Development)
Ángel Clar	(Graphic Design)

Table of Contents

Architectures I

Preservation

Retrieval

Architectures II

Applications

Methodology

Metadata

Evaluation

User Studies

Modeling

Audiovisual Content

Language Technologies

Posters

OpenDLibG: Extending OpenDLib by Exploiting a gLite Grid Infrastructure

Leonardo Candela, Donatella Castelli, Pasquale Pagano, and Manuele Simi

Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo" – CNR
Via G. Moruzzi, 1 - 56124 PISA - Italy
{candela, castelli, pagano, simi}@isti.cnr.it

Abstract. This paper illustrates how an existing digital library system, OpenDLib, has been extended in order to make it able to exploit the storage and processing capability offered by a gLite Grid infrastructure. Thanks to this extension OpenDLib is now able to handle a much wider class of documents than in its original version and, consequently, it can serve a larger class of application domains. In particular, OpenDLib can manage documents that require huge storage capabilities, like particular types of images, videos, and 3D objects, and also create them on-demand as the result of a computational intensive elaboration on a dynamic set of data, although performed with a cheap investment in terms of computing resource.

1 Introduction

In our experience in working with digital libraries (DLs) we have often had to face the problem of resources scalability. Recent technology progresses make it now possible to support DLs where multimedia and multi-type content can be described, searched and retrieved with advanced services that make use of complex automatic tools for feature extraction, classification, summarization, etc. Despite the feasibility of such DLs, the actual use of them is still limited because of the high cost of the computer resources they require. Thus, for example, institutions that need to automatically classify images or 3D objects are forced to afford the cost of large processing capabilities even if this elaboration is only sporadically done. In order to overcome this problem, a couple of years ago we decided to start investigating the use of Grid technologies for supporting an effective handling of these objects. By using the features of the Grid several institutions can share a number of storage and computing resources and use them on-demand, on occasion of their temporary need. This organization allows minimizing the total number of resources required and maximizing their utilization.

Our attempt of exploiting Grid technologies is not isolated. Others are moving in the same direction even if with different objectives. Widely used content repository systems, like DSpace [18] and Fedora [13] as well as DLs, like the Library of Congress, are presently using the SDSC Storage Resource Broker [17] (SRB) as a platform for ensuring preservation and, more in generally, the long term availability of the access to digital information [15, 16].

Chershire3 [14], is an Information Retrieval system that operates both in single processor and in Grid distributed computing environments. A new release of this system

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 1–13, 2006.

capable of processing and indexing also documents stored in the SRB via their inclusion in workflows has been recently designed.

DILIGENT [6] aims at generalizing the notion of sharing proposed by the Grid technologies by creating an infrastructure that connects not only the computing and storage resources but, more generally, all the resources that compose a DL, i.e. archives of information, thesauri, ontologies, tools, etc. By exploiting the functionality provided by DILIGENT, digital libraries will be created on-demand by exploiting the resources connected through this particular infrastructure.

This paper describes how we have extended an existing DL system, OpenDLib [4], in order to make it able to exploit the sharing of storage and processing capabilities offered by a gLite Grid infrastructure[12] for effectively handling new document types. The system resulting from this extension, named *OpenDLibG*, can manage documents that require huge storage capabilities, like particular types of images, videos, and 3D objects, and also create them on-demand as the result of a computational intensive elaboration on a dynamic set of data. The novelty of this system with respect to its predecessor is that, by exploiting the on-demand usage of resources provided by the Grid, it can provide reliable, scalable and high throughput functionality on complex information objects without necessarily large investments on computing resources.

The paper presents the technical solution adopted by highlighting not only the potentialities related to the use of a Grid infrastructure in the particular DL application framework, but also the aspects that have to be carefully considered when designing services that exploit it. The additional features offered by this new version of the OpenDLib system are illustrated by presenting a real application case that has been implemented to concretely evaluate the proposed solution.

The rest of the paper is organized as follows: Section 2 briefly introduces OpenDLib and gLite; Section 3 provides details on the technical solution that has been implemented; Section 4 describes the new functionality OpenDLibG is able to offer and illustrates this functionality by presenting an implemented usage scenario; and finally, Section 5 contains final remarks and plans for further extensions.

2 The Framework

In this section we present a very brief overview of the OpenDLib and gLite technologies by focussing on those aspects that are relevant for describing OpenDLibG.

2.1 OpenDLib

OpenDLib [4] is a digital library system developed at ISTI-CNR. Its is based on a Service-Oriented Architecture that enables the construction of networked DLs hosted by multiple servers belonging to different institutions. Services implementing the DL functionality communicate through a HTTP-based protocol named OLP [5]. These services can be distributed or replicated on more than one server and can be logically organised as in Figure 1.

The *Collective Layer* contains services performing the co-ordination functions (e.g. mutual reconfiguration, distribution and replication handling, work-load distribution) on the services federation. In particular, the *Manager* Service maintains a continually updated status of the DL service instances and disseminates it on request to all the other services that use this information to dispatch message requests to the appropriate service instances. Thanks to this functionality, each service instance does not need to know where the other instances are located and how to discover the appropriate instances to call.

The *DL Components* includes services implementing DL functions. The basic OpenDLib release offers services to support the description, indexing, browsing, retrieval, access, preservation, storage, and virtual organization of documents. In particular, the storage and the dissemination of documents is handled by the *Repository* service.

The *Workflows* provides functionality implemented through workflows, i.e. structured plans of service calls. In particular, this area includes the *Library Manager* which manages and controls the submission, withdrawal and replacement of documents.

The *Presentation* contains services implementing the user front-ends to the other services. It contains a highly customisable *User Interface* and an *OAI-PMH Publisher*.

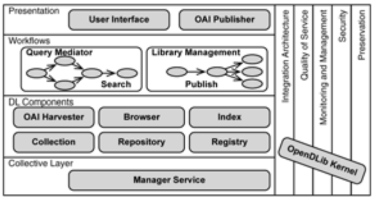


Fig. 1. The OpenDLib Layered Architecture

The *OpenDLib Kernel* supports all the above services by providing mechanisms to ensure the desired quality of service.

These services can be configured by specifying a number of parameters, like metadata and document formats, user profile format, query language, etc. The set of services illustrated above can be extended by including other services that implement additional application-specific functionality.

The OpenDLib services interact by sharing a common information objects model, *DoMDL* [2]. This model can represent a wide variety of information object types with different formats, media, languages and structures. Moreover, it can represent new types of documents that have no physical counterpart, such as composite documents consisting of the slides, video and audio recordings of a lecture, a seminar or a course. It can also maintain multiple editions, versions, and manifestations of the same document, each described by one or more metadata records in different formats. Every manifestation of the digital object can be either locally stored, or retrieved from a remote server and displayed whether at run time or in its remote location. A manifestation can also be a reference to another object manifestation; through this mechanism, data duplication can be avoided.

2.2 gLite

gLite [12] is a Grid middleware recently released by EGEE [7], the largest Grid infrastructure project currently being funded in Europe.

The role of gLite is to hide the heterogeneous nature of both the *computing elements* (CEs), i.e. services representing a computing resource, and *storage elements* (SEs), i.e.

services representing a storage resource, by providing an environment that facilitates and controls their sharing.

The services constituting the gLite software are logically organized as in Figure 2.

The *Job Management Services* is in charge of managing *jobs* and *DAGs*[1]. The core components of this subsystem are the *Computing Element*, the *Workload Manager* (WMS), and the *Logging and Bookkeeping* services. The former represents a computing resource and provides an interface for job submission and control. It is worth noting that the back end of the CE is composed by a set of computational resources managed by a Local Resource Management System (LRMS), e.g. Torque, Condor[2]. The Workload Manager is the subsystem whose main role is to accept requests of job submission and forward them to the appropriate CEs. The Logging and Bookkeeping service is in charge of tracking jobs in terms of *events* - e.g. submitted, running, done - gathered from various WMSs and CEs.

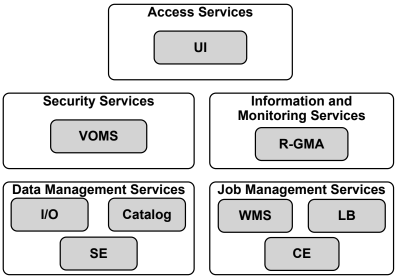


Fig. 2. The gLite Services

The *Data Management Services* is in charge of managing data and file access. gLite assumes that the granularity of data is on file level and that the access is controlled by Access Control Lists. The main services are the *gLite I/O*, the *Storage Element*, and the *Data Catalog*. The former provides a POSIX-like file I/O API, while the Storage Element represents the back end storage resource and can be implemented with various Storage Resource Managers, e.g. dCache[3], DPM[4]. The Data Catalogue allows to perceive the storage capacity of the infrastructure as a single file system.

The *Security Services* is in charge of dealing with authentication, authorization and auditing issues. Actually, the Virtual Organization Membership Service (VOMS) is the main service dealing with these issues. Other aspects are regulated via well known standards and technologies, e.g. X.509 Proxy Certificates [19], Grid Map Files.

The *Information and Monitoring Services* discovers and monitors the resources forming the infrastructure. The main service is the Relational Grid Monitoring Architecture (R-GMA), a registry supporting the adjunction and the removal of data about the resources constituting the infrastructure.

The *Access Services* enables end-users to have access to and use the resources of the infrastructure. Its main component is the User Interface (UI), a suite of clients and APIs enabling users to perform the common user tasks of a gLite based infrastructure, e.g. store and retrieving files, run jobs and monitor on their status.

[1] In gLite terminology jobs are an application that can run on a CE, and DAGs are directed acyclic graphs of dependent jobs.

[2] `http://www.cs.wisc.edu/condor/`

[3] dCache is accessible at `http://www.dcache.org`

[4] DPM information can be found at `http://wiki.gridpp.ac.uk/wiki/Disk_Pool_Manager`

3 OpenDLibG: gLite Integration and Exploitation

The OpenDLib document model is flexible enough to represent a large variety of complex information objects that, if largely employed, could change the way in which research is done. By exploting the functionality built on this model multimedia objects can be composed with table, graphs or images generated by elaborating a large amount of raw data, videos can be mixed with text and geographic information, and so on. Even if the support to this type of complex information objects is theoretically possible with OpenDLib, in practice it turns out to be unrealistic due to the large amount of computing and storage resources that have to be employed to provide performance acceptable by users. Our decision to extend OpenDLib by making it able to exploit the storage and processing capabilities provided by a gLite-compliant infrastructure was mainly motivated by the aim of overcoming this heavy limitation. In the rest of this section we describe the components that we have added, how they have been integrated in the architecture, and the difficulties that we faced in performing this integration.

3.1 The Integrated Architecture

In order to equip OpenDLib with the capabilities required to exploit a gLite-compliant infrastructure we designed the following new services:

- *gLite SE broker*: interfaces OpenDLib services with the pool of SEs made available via the gLite software and optimises their usage.
- *gLite WMS wrapper*: provides OpenDLib services with an interface to the pool of gLite CEs and implements the logic needed to optimize their usage.
- *gLite Identity Provider*: maps the OpenDLib user and service identities onto gLite user identities that are recognized and authorized to use gLite resources.
- *OpenDLib Repository++*: implements an enhanced version of the OpenDLib Repository service. It is equipped with the logic required to manage and optimize the usage of both OpenDLib repositories and gLite SEs as well as to manage novel mechanisms for the dynamic generation of document manifestations.

The architecture of the resulting system is depicted in Figure 3.

Thanks to the extensibility of the OpenDLib application framework the integration of these services has been obtained by only modifying the configuration of some of the existing services without any change in their code. In particular, the OpenDLib Manager Service has been appropriately configured to provide information about the new services and to disseminate new routing rules. These rules enable the OpenDLib UI to interact with instances of the OpenDLib Repository++ service in a completely transparent way for both the submission of, and the access to, documents while the Repository service is only accessed through its new enhanced version.

We explicitly chose to build the enhanced version of the Repository service as an abstraction of the basic version. It does not replace the original service because not all digital libraries require a Grid-based infrastructure. However, this new service maintains all the main characteristics of the basic version and, in particular, it can be replicated and/or distributed in the infrastructure designed to provide the DL application. Finally, the Repository++ can manage a multitude of basic Repository services while a same basic Repository service can accept requests coming from different Repository++.

In the rest of this section we present each of the new services in detail.

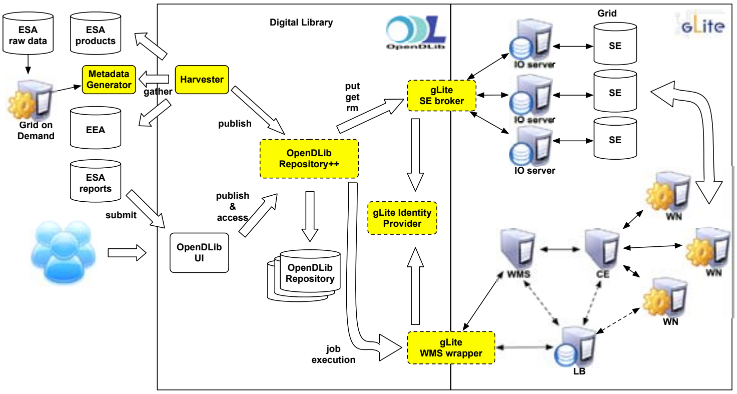


Fig. 3. OpenDLib integrate with gLite: the Architecture

The gLite SE broker. It provides the other OpenDLib services with the capability of using gLite based storage resources. In particular, this service interfaces the *gLite I/O server* to perform the storage and withdrawal of files and the access to them.

In designing this service one of our main concerns was to overcome two problems we have discovered experiencing with the current gLite release: (*i*) the inconsistency between catalog and storage resource management systems and (*ii*) failure in the access or remove operations without notification.

Although the gLite SE broker could not improve the reliability of the requested operations we designed it to: (*i*) monitor its requests, (*ii*) verify the status of the resources after the processing of the operations, (*iii*) repeat file registration in the catalog and/or storage until it is considered correct or unrecoverable, (*iv*) return a valid message reporting the exit status of the operation. The feasibility of this approach was validated by the small resulting delay experimentally measured as well as by real users judgements.

In order to appropriately exploit the great number of resources provided by the infrastructure, the gLite SE broker service was designed to interface more than one I/O server for distributing storage and access operations. In particular, this service can be configured to support three types of storage strategies for distributing files among the I/O servers, namely: (*i*) round-robin, (*ii*) file-type-based, which places the files of a certain type on a predefined set of I/O servers, and (*iii*) priority-based, which is useful to enhance one of the previous strategies with an identified prioritized list of I/O servers ordering the requests to them. It is worth noting that the service can also dynamically rearrange the prioritized list by taking into account performance characteristics, e.g. the time and the number of failures in executing I/O actions.

Inspired by the RAID technology[5] we designed the gLite SE broker to support the *RAID 1* modality that mirrors each file by creating a copy of it on two or more servers. This feature is activated by default but it can be explicitly turned off at configuration time.

The *RAID 0* modality, a.k.a. striped, that splits files on two or more servers and the possibility to select the appropriate modality for each file at the submission time is under investigation.

[5] A Redundant Array of Independent Disks, a.k.a. Redundant Array of Inexpensive Disks.

The gLite WMS wrapper. It provides the other OpenDLib services with the computing power supplied by gLite CEs. In particular, this service offers an interface for managing *jobs* and *DAGs* with an abstraction level higher than that provided by gLite.

The gLite WMS broker has been designed to: *(i)* deal with more than one WMS, (*ii*) monitor the quality of service provided by these WMSs by analyzing the number of managed jobs and the average time of their execution, and, finally, (*iii*) monitor the status of each submitted job querying the *Logging and Bookkeeping* service. As a consequence of the implemented functionality, the gLite WMS service represents a single point of access to the computing capabilities provided by the WMS services and to the monitoring capabilities provided by the LB services. This approach decouples the gLite infrastructure from the OpenDLib federation of services while hiding their characteristics. Moreover, by exploiting the features provided by the OpenDLib application framework, the gLite WMS broker service can be replicated in a number of different service instances, each managing the same set of gLite services, or can be distributed over a number of different service instances, each managing a different pool of gLite services.

In implementing this component we provided both a round-robin and a priority based scheduling strategies to manage the distribution of jobs to WMSs. In particular, the second approach represents an enhancing of the first one because it identifies a priority list of WMSs ordering the requests to them. It is still under investigation the possibility of automatically manipulating this priority in order to take into account performance metrics such as the time and the number of failures.

Finally, we equipped the service with a basic fault tolerance capability in performing job submission tasks that repeats the execution in case of failure.

gLite Identity Provider. The mechanisms that support authentication and authorization in OpenDLib and gLite are different. The two systems have been designed with the aim to satisfy different goals in a completely different usage scenarios: OpenDLib operates in a distributed framework where the participating institutions collaborate and share the same rules and goals under the supervision of a distributed management, while gLite has to work in an environment where policies and access rules are managed differently by the participating institutions. OpenDLib builds its own authentication mechanism on user name and password, while gLite builds it on X.509 Certificates. Moreover, the authorization mechanisms for assigning policies to users are proprietary in OpenDLib while they are based on the Virtual Organization mechanism and Grid Map Files in a gLite based infrastructure. In order to reconcile these authentication and authorization frameworks a service able to map OpenDLib identities on gLite identities was introduced. The main characteristics of this service are:

- it generates the Proxy Certificates [19] that are needed to interact with gLite resources. In order to support this functionality it has to be equipped with the appropriate pool of personal certificates that, obviously, must be stored on a secure device.
- it can be configured to establish the mapping rules for turning OpenDLib identities into gLite identities.

The OpenDLib Repository++. This service was designed to act as a *virtual repository*, capable of the same operations as those required to store and access documents in a traditional OpenDLib DL. In this way the other services of the infrastructure do not need either to be aware of this service's enhanced capabilities nor to be redesigned and re-implemented. Despite the public interface of this service completely resembles the *Repository* interface, its logic is completely different because it does not store any content locally, instead, it relies on the storage facilities provided by both the OpenDLib Repository and the gLite infrastructure via the gLite SE broker.

In designing this component we decided to make the strategy for distributing content on the two kinds of storage systems configurable.

The configuration aspects exploit the DoMDL management functionality allowing any supported manifestation type to be associated with a predefined workflow customising storage, access, and retrieve capabilities. It is thus possible to design and implement the most appropriate processes for each new type of raw data managed by the DL and easily associate it with the manifestation type. In the current version, one workflow to store, access, and retrieve files through the described gLite wrappers has been implemented. For instance, it is possible to configure the Repository++ service in order to maintain all metadata manifestations on a specific OpenDLib Repository instance, a certain manifestation type on another OpenDLib Repository, while raw data and satellite products that are accessed less frequently and require a huge amount of storage can be stored on a SE provided by the gLite based infrastructure. The characteristics of the content to be stored should drive the designer in making the configuration. Usually, manifestations that require to be frequently accessed, or that need to be maintained under the physical control of a specific library institution, should be stored on standard OpenDLib Repository services. On the contrary, content returned by processes, that either is not directly usable by the end-user, or can be freely distributed on third-party storage devices should be stored on gLite SEs.

Cryptography capabilities are under investigation to mitigate the problems mostly related to the copyright management for storing content on third-party devices.

Another important feature added to the enhanced repository is the capability of associating a job or a DAG of jobs with a manifestation. This feature makes it possible to manage new types of document manifestations, i.e., manifestations dynamically generated by running a process at access time. The realisation of such extension has been quite simple in OpenDLib thanks to DoMDL. In fact, DoMDL is able to associate a URI of a specific task with a manifestation. In this case, this task uses the gLite WMS wrapper for executing a process customized with the information identifying the job/DAG to be run together with the appropriate parameters.

An example of the exploitation of this functionality is given in the following section.

4 OpenDLibG in Action: An Environmental DL

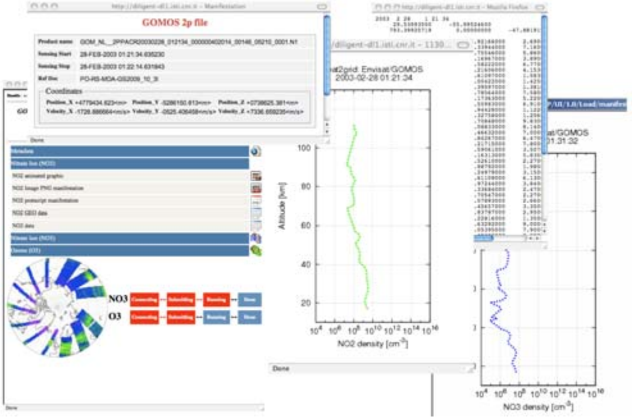
Stimulated by the long discussions we had with members of the European Space Agency (ESA)[6], we decided to experiment the construction of an OpenDLibG DL for supporting the work of the agencies that collaboratively work at the definition of environmental conventions. By exploiting their rich information sources, ranging from raw data sets to

[6] These discussions and the corresponding requirements where raised mainly in the framework of the activities related to the DILIGENT project.

maps and graphs archives, these agencies periodically prepare reports on the status of the environment. Currently, this task is performed by first selecting the relevant information from each of the multiple and heterogeneous sources available, then launching complex processing on large amount of data to obtain graphs, tables and other summary information and, finally, producing the required report by assembling all the different parts together. This process repeated periodically requires a lot of work due to the complexity of interfacing the different sources and tools. Despite the effort spent, the resulting reports do not always met the requirements of the their users since they present to its readers a picture of the environmental status at the time of the production of the report and not at time in which the information reported is accessed and used. To overcome this problem and, more generally, to simplify the generation of the environmental reports we created an OpenDLibG DL prototype.

From the architectural point of view, the OpenDLibG components of this DL are hosted on three servers. The first server is publicly accessible and hosts the User Interface service that allows end-users to easily interact with a human-friendly interface. The second and third servers are protected behind a firewall and host the basic and the extended OpenDLib services respectively.

As far as the Grid infrastructure is concerned, the OpenDLibG environmental DL exploits the DILIGENT gLite infrastructure. This infrastructure consists of five sites located in Pisa (Italy), Rome (Italy) Athens (Greece), Hall in Tyrol (Austria) and Darmstadt (Germany). Each site provides storage and computational capabilities for a total of 41 Processors, 38,72 GB RAM, and 3,28 TB disk space. For the scope of this DL, we decided to exploit only two storage elements based on dCache and other two storage elements based on DPM.



GOMOS 2p file

Product name	GOM_NL__2PPACR20030228_012134_000000402014_00146_05210_0001.N1
Sensing Start	28-FEB-2003 01:21:34.635230
Sensing Stop	28-FEB-2003 01:22:14.631843
Ref Doc	PO-RS-MDA-GS2009_10_3I

Coordinates

Position_X +4779434.623<m> Position_Y -5286150.613<m> Position_Z +0738625.381<m>
Velocity_X -1728.886664<m/s> Velocity_Y -0525.406458<m/s> Velocity_Z +7336.659235<m/s>

Fig. 4. A GOMOS Document

In this experimental environmental DL the Repository service has been configured to manage DoMDL instances that are able to maintain both information objects selected from third-party information sources - whose content is imported/linked in/to the DL - and information objects whose manifestations are generated on-demand using a registered workflow that invokes the gLite WMS Wrapper for executing specific elaborations.

This DL provides the data, the documents, the dynamically generated reports, and any other content and services deemed as relevant with respect to the day-by-day activity of people who have to take decisions on environmental strategies. In particular, the prototype can: (*i*) manage environmental reports, workshops proceedings and other types of documents relevant to the Earth Observation community collected from different and heterogenous information sources; (*ii*) deal with added value satellite products like chlorophyll concentration maps and mosaics; (*iii*) dynamically produce reports related to certain world regions and time periods; (*iv*) use the gLite job management facilities to produce Nitrate and Ozone profiles from satellite products, thus making experiments on the management of such data; (*v*) support search and use of a number of relevant external data, services, and resources concerned with Earth Science, like glossaries, gazetteers, and other digital libraries of interest.

In particular, the above information objects have been obtained by harvesting: (*i*) documents gathered or linked from external information sources like MFSTEP monthly reports[7] and the European Environment Agency[8] reports, briefings, indicators and news, (*ii*) high resolution satellite images both directly acquired and dynamically generated, and (*iii*) level two ENVISAT-GOMOS products[9] containing raw data on ozone, temperature, moisture, NO_2, NO_3, $OClO$, O_3 measures collected by the GOMOS sensor.

Figure 4 shows an example of the novel type of documents that can be managed by this DL. This document is composed by (*i*) a metadata view containing descriptive information like the start and stop sensing dates and the coordinates of the geographical area the document refers to and (*ii*) three products defined via appropriate workflows whose manifestations are generated by running workflows on the Grid infrastructure. These workflows exploit the BEAT2GRID application, provided by the ESA organisation and adapted by us to run on a gLite based infrastructure, and the appropriate operations for gathering from the Grid the raw data to be elaborated, storing on the Grid the obtained products, and linking them as document manifestations. Such workflows generate geolocation information extracted from the raw data; the NO_2/NO_3 image profile information showing the density with respect to the altitude; the NO_2/NO_3 profile information comprising date, time, longitude and latitude of tangent point, longitude and latitude of satellite; the ozone density with respect to the altitude and the ozone density covariance. Each of such products represent a manifestation of a product view. According to the document definition, each product manifestation can be retrieved from the Grid or dynamically generated at access time. To give access to these complex objects a specialised user interface has been designed. It is capable capable to start the product generation process, progressively show the status of the workflow execution and, once products are generated, give access to them.

[7] `http://www.bo.ingv.it/mfstep/`

[8] `http://www.eea.eu.int/`

[9] ENVISAT (ENVIronment SATellite) is an ESA Earth Observation satellite whose purpose is to collect earth observations: it is fitted with 10 sensors ASAR, MERIS, AATSR, RA-2, MWR, DORIS, GOMOS, MIPAS, SCIAMACHY, LRR and other units. Detailed information about the ENVISAT satellite can be found at `http://envisat.esa.int/`

It is worth noting that the BEAT2GRID application is executed in a couple of minutes in a quite normal bi-processor entry-level server. However, standard DL based applications can not provide such functionality to end-users since hundreds of concurrent requests prove the limited scalability of a static infrastructure. OpenDLibG powered by the described gLite based infrastructure, instead, proves to manage tenths of concurrent requests with the same throughput as the single execution on a dedicated server, and to correctly manage a higher number of requests by using queue mechanisms. The same observation holds with respect to the storage capacity. Raw data and intermediate processing results require an huge amount of storage space. Thanks to the Grid technology this space can be obtained on demand by relying on third party institutions while in the case of a standard DL it is needed to equip the DL with such amount of resources even if they are needed only for a limited time period.

This experimentation can be considered the first step in the exploitation of Grid enabled DLs. Moreover it represents a great opportunity for both users and digital library developers to share views and language, to express needs via practical examples, to understand capabilities for future exploitation, to access practical progress, to evaluate opportunities and alternative solutions, to support technical decisions and, last but not least, to develop critical interfaces.

5 Conclusions and Lesson Learned

This paper has described OpenDLibG, a system obtained by extending the OpenDLib digital library system with services exploiting a gLite Grid infrastructure. As a result of this extension OpenDLibG can provide both a more advanced functionality on novel information objects and a better quality of service without requiring a very expensive infrastructure.

We strongly believe that the new information objects described in this paper will play an increasingly important role in the future as they can contribute to revolutionize the way in which many communities perform their activities. In this paper we have shown only an example of the exploitation of such documents, but many others have been suggested us by the many user communities we are in contact with.

The integration of OpenDLib with a Grid infrastructure not only makes it possible to handle the new type of objects but it also supports any functionality whose implementation requires intensive batch computations. For example, periodic complex feature extraction on large document collections or generation and storage of multiple and alternative manifestations for preservation purposes can similarly be supported while maintaining a good quality of service. Our next future plan is to extend the system with novel and distributed algorithms for providing DL functionality relying on the huge amount of computing and storage power provided by the Grid.

While carrying out this experience we have learnt that there are a number of aspects that have to be carefully considered in designing a DL system that exploits a Grid infrastructure. In this framework resources are provided by third-parties and there is a lack of any central control on their availability. These resources can disappear or become unavailable without informing any central authority that, therefore, has no means to prevent it. This problem is made worst by the lack of advanced reservation, i.e. the possibility for a resource user to agree with the resource provider on the availability of a resource for a well established time period and on a given quality of service. This feature is a long term goal in the Grid research area and it is expected that it will be provided in future releases of Grid middleware. This lack has strong implications on the reliability of the Grid resources usage. For example, a document stored on a single SE can be lost if the SE is removed from the Grid by its provider. Appropriate measures have to be taken to reduce the risk induced by this lack. For example, a DL service must be designed in such a way that if the CE running one of its processes disappears, it must be able to recover this malfunction. Other aspects to be carefully taken into account are related to performance. Some of them apply to any Grid infrastructure, while others are more specific and relate to the gLite software and its current release. Perhaps the most important among these aspects is concerned with the communication overhead that arises when using resources spread over the Net. In this context, where resources are SEs and CEs, the decision to ask third-party for the storage or the processing capabilities must be carefully evaluated, i.e. the enhancement obtained must be compared with the overhead needed and the right trade-off among these aspects must be discovered.

Acknowledgments. This work is partially funded by the European Commission in the context of the DILIGENT project, under the 2nd call of FP6 IST priority.

We thank all the DILIGENT partners that have contributed with their suggestions and expertise to this work. Particular thanks go to ESA, which provided us the requirements for live documents, the real data and the specific applications for our experimentation; CERN which supported us in our understanding of the gLite technology; and FhG-IPSI which studied the problem of dynamically defining the structure and visualization of the on-demand documents and which collaborated with CNR for the setting up of the experimentation.

References

1. W. Y. Arms. *Digital Libraries*. The MIT Press, September 2001.
2. L. Candela, D. Castelli, P. Pagano, and M. Simi. From Heterogeneous Information Spaces to Virtual Documents. In *Proceedings of the 8th International Conference on Asian Digital Libraries, ICADL 2005, Bangkok, Thailand, December 2005*, pages 11–22. Springer, 2005.
3. L. Candela, D. Castelli, P. Pagano, and M. Simi. Moving Digital Library Service Systems to the Grid. In *Peer-to-Peer, Grid, and Service-Orientation in Digital Library Architectures*, number 3664 in Lecture Notes in Computer Science, pages 236 – 259. Springer Verlag, 2005.
4. D. Castelli and P. Pagano. A System for Building Expandable Digital Libraries. In *ACM/IEEE Joint Conference on Digital Libraries (JCDL 2003)*, pages 335–345. Springer-Verlag, 2003.
5. D. Castelli and P. Pagano. The OpenDLib Protocol. Technical report, Istituto di Scienza e Tecnologie dell'Informazione “A. Faedo”, CNR, 2004.
6. DILIGENT. A DIgital Library Infrastructure on Grid ENabled Technology. `http://www.diligentproject.org`.
7. EGEE. Enabling Grids for E-science in Europe. `http://public.eu-egee.org`.
8. I. Foster. What is the Grid? A Three Point Checklist. *GRIDtoday*, 1(6), 2002.

9. I. Foster and C. Kesselman, editors. *The Grid: Blueprint for a Future Computing Infrastructure*. Morgan-Kaufmann, 2004.
10. I. Foster, C. Kesselman, J. Nick, and S. Tuecke. The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration. Open Grid Service Infrastructure WG, Global Grid Forum, June 2002.
11. I. Foster, C. Kesselman, and S. Tuecke. The anatomy of the Grid: Enabling scalable virtual organization. *The International Journal of High Performance Computing Applications*, 15(3):200–222, 2001.
12. gLite. Ligthweight Middleware for Grid Computing. `http://glite.web.cern.ch/`.
13. C. Lagoze, S. Payette, E. Shin, and C. Wilper. Fedora: An Architecture for Complex Objects and their Relationships. *Journal of Digital Libraries, Special Issue on Complex Objects*, 2005.
14. R. R. Larson and R. Sanderson. Grid-based digital libraries: Cheshire3 and distributed retrieval. In *JCDL '05: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 112–113, New York, NY, USA, 2005. ACM Press.
15. R. W. Moore and R. Marciano. Building preservation environments. In M. Marlino, T. Sumner, and F. M. S. III, editors, *JCDL*, page 424. ACM, 2005.
16. A. Rajasekar, R. Moore, F. Berman, and B. Schottlaender. From Digital Preservation Lifecycle Management for Multi-media Collections. In *8th International Conference on Asian Digital Libraries, ICADL 2005, Bangkok, Thailand, December 2005*, pages 380–384. Springer, 2005.
17. A. Rajasekar, M. Wan, R. Moore, W. Schroeder, G. Kremenek, A. Jagatheesan, C. Cowart, B. Zhu, S.-Y. Chen, and R. Olschanowsky. Storage Resource Broker - Managing Distributed Data in a Grid. *Computer Society of India Journal, Special Issue on SAN*, 33(4):42–54, October 2003.
18. R. Tansley, M. Bass, and M. Smith. DSpace as an Open Archival Information System: Current Status and Future Directions. In *Proceedings of the 7th European Conference, ECDL 2003, Trondheim, Norway, August 2003*, pages 446–460. Springer-Verlag, 2003.
19. S. Tuecke, V. Welch, D. Engert, L. Pearlman, and M. Thompson. Internet X.509 Public Key Infrastructure (PKI) Proxy Certificate Profile. RFC3820, IETF, The Internet Engineering Task Force, June 2004.

A Peer-to-Peer Architecture for Information Retrieval Across Digital Library Collections*

Ivana Podnar, Toan Luu, Martin Rajman, Fabius Klemm, and Karl Aberer

School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland
{ivana.podnar, vinhtoan.luu, martin.rajman,
fabius.klemm, karl.aberer}@epfl.ch

Abstract. Peer-to-peer networks have been identified as promising architectural concept for developing search scenarios across digital library collections. Digital libraries typically offer sophisticated search over their local content, however, search methods involving a network of such stand-alone components are currently quite limited. We present an architecture for highly-efficient search over digital library collections based on structured P2P networks. As the standard single-term indexing strategy faces significant scalability limitations in distributed environments, we propose a novel indexing strategy–*key-based indexing*. The keys are term sets that appear in a restricted number of collection documents. Thus, they are discriminative with respect to the global document collection, and ensure scalable search costs. Moreover, key-based indexing computes posting list joins during indexing time, which significantly improves query performance. As search efficient solutions usually imply costly indexing procedures, we present experimental results that show acceptable indexing costs while the retrieval performance is comparable to the standard centralized solutions with TF-IDF ranking.

1 Introduction

Research in the area of information retrieval has largely been motivated by the growth of digital content provided by digital libraries (DLs). Today DLs offer sophisticated retrieval features, however, search methods are typically bound to a single stand-alone library. Recently, peer-to-peer (P2P) networks have been identified as promising architectural concepts for integrating search facilities across DL collections [1, 2]. P2P overlays are self-organizing systems for decentralized data management in distributed environments. They can be seen as a common media for 'advertising' DL contents e.g. to specialists in a particular area, or to the broader public. We argue that a wide range of topic and genre specific P2P search engines can facilitate larger visibility of existing DLs while providing guaranties for objective search and ranking performance. Note

* The work presented in this paper was carried out in the framework of the EPFL Center for Global Computing and supported by the Swiss National Funding Agency OFES as part of the European FP 6 STREP project ALVIS (002068).

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 14–25, 2006.

that P2P networks cannot be centrally controlled: Peers are located in various domains requiring minimal in place infrastructure and maintenance.

Full-text P2P search is currently an active research area as existing P2P solutions still do not meet the requirements of relevance-based retrieval. It is a challenging problem since search engines traditionally rely on central coordination, while P2P is inherently decentralized. For example, global document collection statistics are not readily available in P2P environments, and naïve broadcast solutions for acquiring such statistics induce huge network traffic. In fact, scalability issues and potentially high bandwidth consumption are one of the major obstacles for large-scale full-text P2P search [3].

In this paper we present an integrated architecture for information retrieval over textual DL collections. We assume DLs are cooperative and provide an index of a representative sample of their collections, or supply documents they want to make searchable through a P2P engine. In this way DLs can choose the content that becomes globally available, which naturally resolves the problems related to restricted crawler access. The architecture accommodates distributed indexing, search, retrieval, and ranking over structured P2P networks by means of a common global inverted index, and serves as a blueprint for our prototype system ALVIS PEERS, a full-text search engine designed to offer highly-efficient search with retrieval quality comparable to centralized solutions. It is the result of our research efforts within the project ALVIS[1] that aims at building an open-source semantic search engine with P2P and topic specific technology at its core [4].

We propose a novel indexing scheme and design a distributed algorithm for maintaining the global index in structured P2P networks. Our engine indexes *keys*—terms and term sets appearing in a restricted number of global collection documents—while keeping indexing at document granularity. Indexed keys are rare and discriminative with respect to a global document collection. They represent selective queries readily retrievable from the global P2P index, while search costs are significantly reduced due to limited posting list size. As our engine provides highly-efficient search over a global P2P network, the indexing procedure is costly. However, since DL collections are rather static, it is appropriate to invest resources into the indexing procedure and benefit largely from the search performance. We will show experimentally that, as we carefully choose keys, the key indexing costs remain acceptable. The number of indexed keys per peer is nearly constant for large document collections, as well as the average posting list size when we keep the number of documents per peer constant and increase the global collection by adding new peers. The bandwidth consumption during retrieval is substantially smaller compared to single-term indexing, while the observed retrieval quality (top-k precision) is comparable to the standard centralized solutions with TF-IDF ranking. In contrast to the majority of published experimental results that rely on simulations, our experiments have been performed using a fully fledged prototype system built on top of the P-Grid P2P platform[2].

The paper is structured as follows. Section 2 reviews the characteristics of P2P networks in the context of full-text search, while Section 3 presents our novel key-based indexing strategy. Section 4 specifies the integrated architecture for P2P full-text search

[1] http://www.alvis.info/

[2] http://www.p-grid.org/

and defines a distributed algorithm for building the key index. Experimental results investigating indexing costs and retrieval performance are presented in Sect. 5. Section 6 briefly covers related work, and we conclude the paper in Section 7.

2 Unstructured vs. Structured P2P

There are two main categories of P2P systems, unstructured and structured. In unstructured systems peers broadcast search requests in the network, which works well if used to search for popular highly-replicated content. However, broadcast performs poorly if used to search for rare items as many messages are sent through the network. More advanced approaches restrict the amount of query messages by using random walks [5] or special routing indexes, which maintain content models of neighboring peers in order to determine routing paths for a query [6]. The second class is structured P2P, also called structured overlay networks or distributed hash tables (DHT) [7, 8, 9]. In structured P2P, each peer is responsible for a subset of identifiers id in a common identifier space. Multiple peers may be responsible for the same identifier space to achieve higher reliability. All peers use an overlay routing protocol to forward messages for which they are not responsible. To allow efficient routing, most DHTs maintain routing tables of size $O(log(N))$ where N is the number of peers in the network. Starting at any peer in the network, a message with any destination id can be routed in $O(log(N))$ overlay hops to the peer responsible for id. Structured P2P overlay networks therefore exhibit much lower bandwidth consumption for search compared to unstructured networks. However, they are limited to exact-match key search. Please refer to [10] for a comprehensive analysis of generic P2P properties.

There are two architectural concepts for designing P2P search engines in the area of information retrieval: a) local indexes in unstructured/hierarchical P2P networks, and b) global index in structured P2P networks. The first strategy [6] divides documents over the peer network, and each peer maintains the index of its local document collection. Such indexes are in principle independent, and a query is broadcasted to all the peers in unstructured networks generating an enormous number of messages. To limit the query traffic, the query can be answered at two levels, the peer and document level: The first step locates a group of peers with potentially relevant document collections, while in the second step the query is submitted to the peers, which then return answers by querying their local indexes. The answers are subsequently merged to produce a single ranked hit list. The second strategy [11] distributes the global document index over a structured P2P network. Each peer is responsible for a part of the global vocabulary and their associated posting lists. A posting list consists of references to the documents that contain the associated index term. Queries are processed by retrieving posting lists of the query terms from the P2P network. Our approach is based on the second strategy.

3 Our Approach: Indexing Rare Keys

The key idea of our indexing strategy is to limit the posting list size of the global P2P index to a constant predefined value and extend the index vocabulary to improve retrieval effectiveness. Fig. 1 compares our *rare-key indexing strategy* to the standard single-term indexing approach. It is visible that we trade in an increased index vocabulary for the limited posting list size. As posting lists are extremely large for a single-term index, the process of joining them at query time generates unacceptable network traffic, which makes this approach practically unfeasible. On the other contrary, rare-key indexing offers highly-efficient query performance as we limit the posting list size according to network characteristics and intersect posting lists at indexing time.

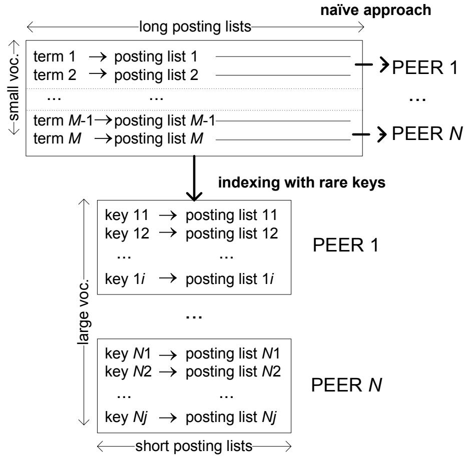


Fig. 1. The basic idea of indexing with rare keys

Let D be a global document collection, and T its single-term vocabulary. A key $k \in K$ consists of a set of terms $\{t_1, t_2, \ldots, t_s\}$, $t_i \in T$, appearing within the same document $d \in D$. The number of terms comprising a key is bounded, i.e. $1 \leq s \leq s_{max}$. The quality of a key k for a given document d with respect to indexing adequacy is determined by its *discriminative power*. To be *discriminative*, a key k must be as specific as possible with respect to d and the corresponding document collection D [12]. We categorize a key on the basis of its *global document frequency* (DF), and define a threshold DF_{max} to divide the set of keys K into two disjoint classes, a set of rare and frequent keys. If a key k appears in more than DF_{max} documents, i.e. $DF(k) > DF_{max}$, the key is *frequent*, and has low discriminative power. Otherwise, k is *rare* and specific with respect to the document collection.

Although the size of the key vocabulary is bounded for a bounded collection size of limited size documents, there are many term combinations that form potential rare keys and special filtering methods are needed to reduce the key vocabulary to a practically manageable size. We currently use the proximity and redundancy filter to produce *highly-discriminative keys* (HDKs) indexed by our search engine. *Proximity filter* uses textual context to reduce the size of the rare key vocabulary and retains keys built of terms appearing in the same textual context—a document window of predefined size w—because words appearing close in documents are good candidates to appear together in a query. The analysis presented in [13] reports the importance of text passages that are more responsive to particular user needs than the full document. *Redundancy filter* removes supersets of rare keys from the vocabulary as such keys are redundant and only increase the vocabulary size without improving retrieval performance. Therefore, all properly contained term subsets in rare keys are frequent, and we call such keys *intrinsically rare* (i-rare) keys. Proximity filtering strongly depends on the window size and document characteristics. Although it seems intuitive that it would remove most keys, our experiments show the great importance of the redundancy filter which removes many keys after proximity filtering (e.g. 83% of 2-term and 99% of 3-term keys). By applying both the proximity and redundancy filter to rare keys, we obtain a significantly smaller set of HDKs compared to the theoretical value, as reported in Section 5.

As our engine indexes keys, it is essential to map queries to keys for an effective retrieval performance. We will now discuss the problem of finding, given a query $Q = \{t_1, t_2, \ldots, t_q\}, t_i \in T$, the corresponding relevant keys in the HDK index. A perfect situation occurs when $\{t_1, t_2, \ldots, t_q\}$ is an HDK, in other words, a user has posed a good discriminative query for the indexed document collection: The posting list is readily available and is simply retrieved from the global index. However, this may not happen with all user queries. Therefore, we use terms and term sets from Q to form potential HDKs. We extract all the subsets of $s_{max}, (s_{max} - 1), \ldots, 1$ terms from the query Q to retrieve the posting lists associated with the corresponding keys, and provide a union of retrieved posting lists as an answer to Q. In fact, we first check s_{max}-term combinations, and if all of them retrieve posting list, we stop the procedure because there will be no $(s_{max} - 1)$-term HDKs. For example, for a query $Q = \{t_1, t_2, t_3\}$ and $s_{max} = 2$, possible 2-term keys are $\{t_1, t_2\}$, $\{t_1, t_3\}$, and $\{t_2, t_3\}$. If we retrieve postings for $\{t_1, t_2\}$ and $\{t_1, t_3\}$, there is no need to check whether $\{t_1\}$, $\{t_2\}$, or $\{t_3\}$ are indexed because i-rare keys cannot be subsets of other i-rare keys. If we retrieve a posting only for $\{t_1, t_2\}$, we still need to check $\{t_3\}$, as it may be an HDK. A similar query mapping principle has recently been proposed for structuring user queries into smaller maximal term sets [14].

However, users may still pose queries containing only frequent keys, or some query terms may not be covered by HDKs. A valid option is to notify a user that his/her query in non-discriminative with respect to the document collection, and provide support for refining the query. We have also devised two other possible strategies to improve the retrieval performance in such cases: The first strategy uses *distributional semantics* [15] to find semantically similar terms to query terms, while the second strategy indexes k-best documents for frequent keys, as the size of the frequent key vocabulary is less than 1% of the HDK size. We leave further analysis of the two strategies for future work.

4 Architecture

We assume an environment comprising a set of M independent DLs hosting local document collections and willing to make a part of their collections searchable through a global distributed index. Each DL is a standalone component that can index and search its local document collection, and therefore provide (a part of) its *local single-term index* as a contribution to the global index. A structured P2P network with N peers is available to share a *global index*, and offer efficient search over the global collection composed of documents contributed by M DLs.

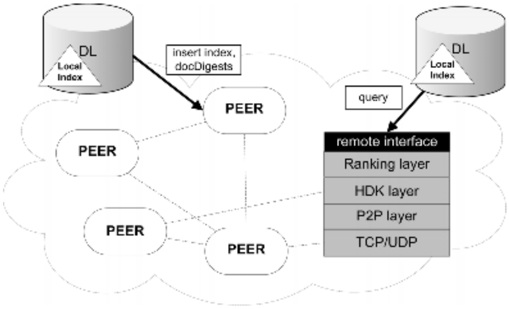


Fig. 2. An overview of the P2P architecture for digital libraries

The high-level architecture of our P2P search engine is presented in Fig. 2. DLs interact with peers to *submit an index* and to *send a query* to the engine. A peer can be regarded as an entry point to a distributed index, and a P2P network as a scalable and efficient media for sharing information about DL content. The architecture is layered to enable clean separation of different concepts related to P2P networks, document and content modeling, and the applied retrieval model [16]. As the global index is key-based, the system is decomposed into the following four layers: 1) transport layer (TCP/UDP) providing the means for host communication; 2) P2P layer building a distributed hash table and storing global index entries; 3) HDK layer for building a key vocabulary and corresponding posting lists, and mapping queries to keys; and 4) Ranking layer that implements distributed document ranking.

Each peer incorporates a *local and global system view*. The HDK layer focuses on the local view and builds the key index from a received single-term index for a DL's local collection. The received single-term index must contain a positional index needed for key computation, and may provide DL's relevance scores for (term, document) pairs. The P2P layer provides a global system view by maintaining the global key index with information about rare and frequent keys. Global index entries have the following structure $\{k, DF(k), PeerList(k), Posting(k)\}$, where $DF(k)$ is the key's global document frequency, $PeerList(k)$ is the list of peers that have reported local document frequencies $df(k)$, and $Posting(k)$ is the k's global posting list. The $Posting(k)$ is $null$ in case k is frequent.

4.1 Distributed Indexing

The indexing process is triggered when a DL inserts a single-term index or document collection into the P2P search engine. Since the indexing process is computationally intensive, peers share computational load and build the HDK vocabulary in parallel. Each peer creates HDKs from the received index, inserts local document frequencies for HDKs it considers locally i-rare or frequent, and subsequently inserts posting lists for globally i-rare keys into the P2P overlay. The P2P layer stores posting lists for globally i-rare keys, maintains the global key vocabulary with global DFs, and notifies the HDK layer when i-rare keys become frequent due to addition of new documents.

Algorithm 1 defines the process of computing HDKs locally by peer P_i at its HDK layer. It is performed in levels by computing single-term, 2-term, ..., s_{max}-term keys. The peer stores a set of potentially i-rare keys in K_{ir}, and globally frequent keys in K_{freq}. Note that a locally frequent key is also globally frequent, but each locally rare key may become globally frequent. The P2P overlay is aware when a key becomes frequent, and notifies interested peers from the $PeerList(k)$.

The algorithm starts by inserting local document frequencies for the single-term vocabulary T_i and classifying terms as frequent or rare. Note that a peer is notified when its locally rare keys become globally frequent, which depends on the HDK computation process performed by other peers. Next, P_i re-checks single-term DFs, and inserts posting lists for the rare ones into the P2P overlay. The approach is tolerant to erroneous insertions of posting lists for frequent keys: The P2P overlay disregards the received posting list, updates the global document frequency of a key, and notifies a peer that the key is frequent.

For determining multi-term i-rare keys, the algorithm uses term locations from the received single-term index. A potential term combination needs to appear within a predefined window, next the redundancy property is checked, and if a key passes both filters, it is an HDK candidate. It's global frequency is updated in the P2P overlay, but the HDK layer at this point updates its posting list only locally. The global posting list will be updated subsequently in case the key was not reported globally frequent by the P2P layer.

4.2 Distributed Retrieval

The query and retrieval scenario involves all four architectural layers. A query is submitted through a peer's remote interface to the HDK layer which maps query terms to HDKs as discussed in Section 3. The HDK layer retrieves posting lists associated with relevant HDKs from the global P2P index. The received posting lists are merged, and submitted to the ranking layer. The ranking layer ranks documents, and must be designed to provide relevance scores with the minimal network usage. There are a number of ranking techniques the proposed architecture can accommodate, but here we only sketch an approach using content-based ranking since distributed ranking is outside the scope of this paper.

As the P2P index maintains global DFs for all frequent and rare terms, DFs for the vocabulary T are readily available in the index and may be retrieved to be used for ranking. Term frequencies are local document-related values that are also used for computing content-based relevance scores. As DLs provide either a single-term index or original documents when initiating the indexing procedure, the indexing peer can use them to extract/compute document-related term statistics. Consequently, we can rank an answer set using a relevance ranking scheme that relies on global document frequencies and term frequencies, without knowing the total global document size, as this parameter is typically used to normalize the scores.

Algorithm 1. Computing HDKs at peer P_i

```
for $s = 1$ to $s_{max}$ do
    $K^s_{ir} \leftarrow \emptyset$
    $K^s_{freq}(s) \leftarrow \emptyset$
    if $s = 1$ then
        /* process single-term keys */
        for all $t_k \in T_i$ do
            P2P.updateDF($key$)
            if $df(t_k) \leq DF_{max}$ then
                $K^s_{ir} \leftarrow K^s_{ir}(s) \cup t_k$
            else
                $K^s_{freq} \leftarrow K^s_{freq} \cup t_k$
            end if
        end for
    else
        /* generate new keys from frequent keys*/
        for all $key = (t_{k_1}, \ldots, t_{k_{s-1}}) \in K^{s-1}_{freq}$ do
            /* process each document in the key posting list to create a set of potential term
            combinations */
            for all $d_j \in$ localPostingList($key$) do
                for all $t_{k_s} \in$ windowOf($key$) do
                    $newKey$ = concat($key, t_{k_s}$)
                    if checkRedundancy($newKey$) then
                        $K^s_{ir} \leftarrow K^s_{ir} \cup newKey$
                        P2P.updateDF($newKey$)
                        updateLocalPostingList($newKey, d_j$)
                    end if
                end for
            end for
        end for
    end if
    /* update global key frequency and insert posting list for i-rare*/
    for all $key \in (K^s_{ir} \cup K^s_{freq})$ do
        if $DF(key) > DF_{max}$ then
            /* $key$ is globally frequent */
            $K^s_{ir} \leftarrow K^s_{ir} \backslash key$
            $K^s_{freq}(s) \leftarrow K^s_{freq} \cup key$
        else
            P2P.insertPostingList($key$)
        end if
    end for
end for
```

5 Experimental Evaluation

Experimental setup. The experiments were carried out using a subset of news articles from the Reuters corpus[3]. The documents in our test collection contain between 70 and

[3] http://about.reuters.com/researchandstandards/corpus/

3000 words, while the average number of terms in a document is 170, and the average number of unique terms is 102. To simulate the evolution of a P2P system, i.e. peers joining the network and increasing the document collection, we started the experiment with 2 peers, and added additional 2 peers at each new experimental run. Each peer contributes with 5000 documents to the global collection, and computes HDKs for its local documents. Therefore, the initial global document collection for 2 peers is 10,000 documents, and it is augmented by the new 10,000 documents at each experimental run. The maximum number of peers is 16 hosting in total the global collection of 80,000 documents. The experiments were performed on our campus intranet. Each peer runs on a Linux RedHat PC with 1GB of main memory connected by a 100 Mbit Ethernet. The prototype system is implemented in Java.

Performance analysis. Experiments investigate the number of keys generated by our HDK algorithm, and the resulting average posting list size maintained by the P2P network. All documents were pre-processed: First we removed 250 common English stop words and applied the Porter stemmer, and then we removed 100 extremely frequent terms (e.g. the term ‘reuters’ appears in all the news). The DF_{max} is set to 250 and 500, s_{max} is 3, and $w = 20$ for the proximity filter.

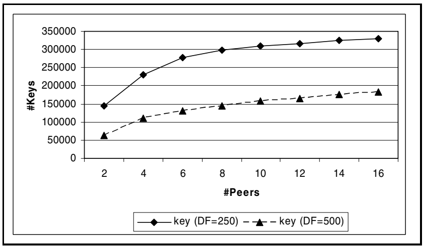


Fig. 3. Average HDK vocabulary per peer

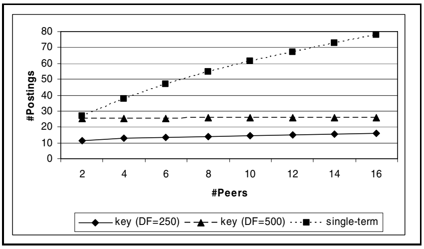


Fig. 4. Average posting list size

Figure 3 shows the total number of HDKs stored per peer for $DF_{max} = 250$ and $DF_{max} = 500$. As expected, an increased value of DF_{max} results in decreased key vocabulary. Both experimentally obtained result sequences exhibit a logarithmic growth and are expected to converge to a constant value because the number of generated term combinations is limited by the proximity window and the total key vocabulary size grows linearly with the global collection size for large collections. The number of keys is quite large compared to the single-term vocabulary, but we expect to benefit from the query performance.

Figure 4 shows the average posting list size for the HDK and single-term indexing. As the average posting list size for HDK indexing remains constant, the expected bandwidth consumption is significantly smaller than for the single-term index exhibiting a linear increase.

For the retrieval performance evaluations, we have created a total of 200 queries by randomly choosing 2 to 3 terms from the news titles. Because of the lack of relevance judgments for our query set, we compared the retrieval performance to a centralized baseline[4] by indexing the collection using both single-term and HDK indexing with deferent DF_{max} values (200, 250, 500). Then for each query we compared the top 20 documents retrieved by our prototype and by the baseline, both hit lists have been ranked using TF-IDF. We are interested in the high-end ranking as typical users are often interested only in the top 20 results. Two metrics are used to compare the result sets: the first one is the overlap between our system and the centralized baseline, and the second one is the average number of posting lists transmitted during retrieval.

Table 1. Retrieval quality of HDK indexing compared to the centralized TF-IDF system

	Overlap ratio on top20	Transmitted postings
single-term (TF-IDF)	100 %	3052.675
HDK, $DF_{max} = 500$	94.34%	232.925 (7.63%)
HDK, $DF_{max} = 250$	85.88%	96.91 (3.17%)
HDK, $DF_{max} = 200$	83.06%	75.37 (2.47%)

Table 1 presents our findings related to retrieval performance for the collection of 30,000 documents over 6 peers. The results show an extreme reduction of the average number of transmitted postings per query of the HDK compared to a naïve P2P approach with single-term indexing which compensates for the increased indexing costs. The results show acceptable retrieval performance of the HDK approach. As expected, the retrieval performance is better for larger DF_{max} as we are getting closer to the single-term indexing, but the average number of transmitted postings also increases, although it is still significantly smaller compared to the single-term case.

6 Related Work

Full-text P2P search is investigated in two overlapping domains: DLs and the Web. There is an ongoing debate on the feasibility of P2P Web search for scalability reasons. In [3] it is shown that the naïve use of unstructured or structured overlay networks is practically infeasible for the Web, since the generated traffic required for indexing and search exceeds the available Internet capacity. Thus different schemes have been devised to make P2P Web search feasible. Several approaches target at a term-to-peer indexing strategy, where the unit of indexing are peers rather than individual documents: PlanetP [17] gossips compressed information about peers' collections in an unstructured P2P network, while MINERVA [18] maintains a global index with peer collection statistics in a structured P2P overlay to facilitate the peer selection process.

As DLs represent only a small fraction of the entire Web space, the feasibility of full-text P2P search across DL collections is not in question. Hierarchical solutions have been investigated for federated search where a backbone P2P network maintains a directory service to route queries to peers with relevant content [6, 1]. A recently proposed solution uses collection-wide statistics to update routing indexes dynamically at query time, and reports low traffic overheads for the Zipf-distribution queries after the

[4] Terrier search engine, http://ir.dcs.gla.ac.uk/terrier/

initial ‘learning phase’ [19]. These solutions are orthogonal to our approach since they are designed for unstructured P2P networks with the low-cost indexing schemes, while the processing and major network traffic is generated during the query phase. Our technique is costly in terms of indexing, however, it offers highly-efficient and responsive querying performance. It is comparable to solutions for distributed top-k retrieval that aim at minimizing query costs by transmitting a limited number of postings [19, 20]. However, the major difference is our novel indexing strategy. The HDK approach is not the only indexing strategy that uses term sets as indexing features. The set-based model [21] indexes term sets occurring in queries, and exploits term correlations to reduce the number of indexed term sets. The authors report significant gains in terms of retrieval precision and average query processing time, while the increased index processing time is acceptable. In contrast to our indexing scheme, the set-based model has been used to index frequent term sets and is designed for a centralized setting.

7 Conclusion

We have presented a P2P architecture for information retrieval across digital library collections. It relies on a novel indexing strategy that indexes rare terms and term sets to limit the bandwidth consumption during querying and enable scalable and highly-efficient search performance. As a proof of concept, we have implemented a prototype system following the presented architectural design, and performed experiments to investigate query performance and indexing costs. Our experiments have demonstrated significant benefits of the HDK approach in terms of reduced networking costs and the feasibility of the proposed indexing strategy for P2P environments. Our future work will further investigate techniques for reducing the cost of the proposed indexing strategy, e.g., by using query statistics, or query-driven indexing. We will perform experiments with larger and various document collections, and increased size of the peer network to confirm existing positive results related to both the networking costs and retrieval performance.

References

1. Lu, J., Callan, J.: Federated search of text-based digital libraries in hierarchical peer-to-peer networks. In: Advances in Information Retrieval, 27th European Conference on IR Research (ECIR). (2005) 52–66
2. Balke, W.T., Nejdl, W., Siberski, W., Thaden, U.: DL Meets P2P - Distributed Document Retrieval Based on Classification and Content. In: 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL). (2005) 379–390
3. Li, J., Loo, B., Hellerstein, J., Kaashoek, F., Karger, D., Morris, R.: The feasibility of peer-to-peer web indexing and search. In: Peer-to-Peer Systems II: 2nd International Workshop on Peer-to-Peer Systems (IPTPS). (2003) 207–215
4. Buntine, W., Aberer, K., Podnar, I., Rajman, M.: Opportunities from open source search. In: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence. (2005) 2–8
5. Lv, Q., Cao, P., Cohen, E., Li, K., Shenker, S.: Search and replication in unstructured peer-to-peer networks. In: 16th International Conference on Supercomputing. (2002) 84–95

6. Lu, J., Callan, J.: Content-based retrieval in hybrid peer-to-peer networks. In: Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM). (2003) 199–206
7. Ratnasamy, S., Francis, P., Handley, M., Karp, R., Shenker, S.: A scalable content-addressable network. In: SIGCOMM '01. (2001) 161–172
8. Stoica, I., Morris, R., Karger, D., Kaashoek, M.F., Balakrishnan, H.: Chord: A scalable peer-to-peer lookup service for internet applications. In: SIGCOMM '01. (2001) 149–160
9. Aberer, K.: P-Grid: A self-organizing access structure for P2P information systems. In: CooplS '01: Proceedings of the 9th International Conference on Cooperative Information Systems. (2001) 179–194
10. Aberer, K., Alima, L.O., Ghodsi, A., Girdzijauskas, S., Haridi, S., Hauswirth, M.: The Essence of P2P: A Reference Architecture for Overlay Networks. In: Fifth IEEE International Conference on Peer-to-Peer Computing. (2005) 11–20
11. Reynolds, P., Vahdat, A.: Efficient Peer-to-Peer Keyword Searching. Middleware03 (2003)
12. Salton, G., Yang, C.: On the specification of term values in automatic indexing. Journal of Documentation **4** (1973) 351–372
13. Salton, G., Allan, J., Buckley, C.: Approaches to Passage Retrieval in Full Text Information Systems. In: SIGIR'93. (1993) 49–58
14. Pôssas, B., Ziviani, N., Ribeiro-Neto, B., Wagner Meira, J.: Maximal termsets as a query structuring mechanism. In: CIKM '05. (2005) 287–288
15. Rajman, M., Bonnet, A.: Corpora-Base Linguistics: New Tools for Natural Language Processing. 1st Annual Conference of Association for Global Strategic Information (1992)
16. Aberer, K., Klemm, F., Rajman, M., Wu, J.: An Architecture for Peer-to-Peer Information Retrieval. In: SIGIR'04, Workshop on Peer-to-Peer Information Retrieval. (2004)
17. Cuenca-Acuna, F.M., Peery, C., Martin, R.P., Nguyen, T.D.: PlanetP: Using Gossiping to Build Content Addressable Peer-to-Peer Information Sharing Communities. In: 12th IEEE International Symposium on High Performance Distributed Computing (HPDC-12), IEEE Press (2003) 236–246
18. Bender, M., Michel, S., Triantafillou, P., Weikum, G., Zimmer, C.: Improving collection selection with overlap awareness in P2P search engines. In: SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. (2005) 67–74
19. Balke, W., Nejdl, W., Siberski, W., Thaden, U.: Progressive distributed top-k retrieval in peer-to-peer networks. In: Proceedings of the 21st International Conference on Data Engineering (ICDE 2005). (2005) 174–185
20. Michel, S., Triantafillou, P., Weikum, G.: KLEE: a framework for distributed top-k query algorithms. In: VLDB '05. (2005) 637–648
21. Pôssas, B., Ziviani, N., Wagner Meira, J., Ribeiro-Neto, B.: Set-based vector model: An efficient approach for correlation-based ranking. ACM Trans. Inf. Syst. **23** (2005) 397–429

Scalable Semantic Overlay Generation for P2P-Based Digital Libraries

Christos Doulkeridis[1], Kjetil Nørvåg[2], and Michalis Vazirgiannis[1]

[1] Dept. of Informatics, AUEB, Athens, Greece
{cdoulk, mvazirg}@aueb.gr
[2] Dept. of Computer Science, NTNU, Trondheim, Norway
Kjetil.Norvag@idi.ntnu.no

Abstract. The advent of digital libraries along with the tremendous growth of digital content call for distributed and scalable approaches for managing vast data collections. Peer-to-peer (P2P) networks emerge as a promising solution to delve with these challenges. However, the lack of global content/topology knowledge in an unstructured P2P system demands unsupervised methods for content organization and necessitates efficient and high quality search mechanisms. Towards this end, Semantic Overlay Networks (SONs) have been proposed in the literature, and in this paper, an unsupervised method for decentralized and distributed generation of SONs, called DESENT, is proposed. We prove the feasibility of our approach through analytical cost models and we show through simulations that, when compared to flooding, our approach improves recall by more than 3-10 times, depending on the network topology.

1 Introduction

The advent of digital libraries along with the tremendous growth of digital content call for distributed and scalable approaches for managing vast data collections. Future digital libraries will enable citizens to access knowledge any time/where, in a friendly, multi-modal, efficient and effective way. Reaching this vision requires development of new approaches that will significantly reform the current form of digital libraries. Key issues in this process are [9]: the system architecture and the information access means. With respect to system architecture, peer-to-peer (P2P) is identified as a topic of primary interest, as P2P architectures allow for loosely-coupled integration of information services and sharing of information/knowledge [1,6,11].

In this paper, we present a scalable approach to P2P document sharing and retrieval. Because scalability and support for semantics can be difficult in structured P2P systems based on DHTs, we instead base our approach on *unstructured P2P networks*. Such systems, in their basic form, suffer very high search costs, in terms of both consumed bandwidth and latency, so in order to be useful for real applications, more sophisticated search mechanisms are required. We solve this problem by employing *semantic overlay networks* (SONs) [5], where peers containing related information are connected together in separate overlay networks. If SONs have been created, queries can be forwarded to only those peers containing documents that satisfy the constraints of the query context, for example based on topic, user profiles or features extracted from previous queries.

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 26–38, 2006.

One of the problems of SONs is the actual construction of these overlays, because in a P2P context there is a lack of knowledge of both global content and network topology. In a P2P architecture, each peer is initially aware only of its neighbors and their content. Thus, finding other peers with similar contents, in order to form a SON, becomes a tedious problem. This contrasts to a centralized approach, where all content is accessible to a central authority, and clustering becomes a trivial problem, in the sense that only the clustering algorithm (and its input parameter values) determines the quality of the results.

The contribution of this paper is a *distributed* and *decentralized* method for hierarchical SON construction (DESENT) that provides an efficient mechanism for search in unstructured P2P networks. Our strategy for creating SONs is based on clustering peers based on their content similarity. This is achieved by a recursive process that starts on the individual peers. Through applying a clustering algorithm on the documents stored at the peer, one or more feature vectors are created for each peer, essentially one for each topic a peer covers. Then representative peers, each responsible for a number of peers in a *zone* are selected. These peers, henceforth called *initiators*, will collect the feature vectors from the members of the zone and use these as basis for the next level of clustering. This process is applied recursively, until we have a number of feature vectors covering all available documents.

The organization of the rest of this paper is as follows. In Section 2, we give an overview of related work. In Section 3, we present our method for creating SONs that can be used in the search process (Section 4). In Section 5, we use analytical cost models to study the cost and the time required for overlay creation, while, in Section 6, we present the simulation results. Finally, in Section 7, we conclude the paper.

2 Related Work

Several techniques have been proposed that can improve search in unstructured P2P systems [2,8], including techniques for improved routing that give a direction towards the requested document, like routing indices [4], and connectivity-based clustering that creates topological clusters that can be used as starting points for flooding [12]. An approach to improve some of the problems of Gnutella-like systems [2], is to use a super-peer architecture [15], which can be also used to realize a hierarchical summary index, as described in [13].

The concept of semantic overlay networks (SONs) [5] is about directing searches only to a specific subset of peers with content relevant to the query. The advantage of this approach is that it reduces the flooding cost in the case of unstructured systems. Crespo and Garcia-Molina [5] essentially base their approach on partly pre-classified documents that only consist of information about the song contained in a particular file. Also they do not provide any other algorithm for searching, other than flooding. In order to be useful in a large system, unsupervised and decentralized creation of SONs is necessary, as well as efficient routing to the appropriate SON(s). The DESENT approach described in our paper solves these issues.

Although several papers describe how to use SON-like structures for P2P content search [3,10], little work exists on the issue of how to actually create SONs in an unsupervised, decentralized and distributed way in unstructured networks. Distributed clustering in itself is considered a challenge demanding for efficient and effective solutions. In [14], a P2P architecture where nodes are logically organized into a fixed number of clusters is presented. The main focus of the paper is fairness with respect to the load of individual nodes. In contrast to our approach, the allocation of documents to clusters is done by classification, it is not unsupervised, and clusters are not hierarchical. We believe that current research in P2P digital libraries [1,6,11] can benefit from the merits of our approach.

3 Overlay Network Creation

In this section, we describe SON generation, assuming peers storing digital content and being connected in an unstructured P2P network. Each peer represents a digital library node and in this paper we focus on peers that store documents, though other data representations can also be supported. The approach is based on creating local zones of peers, forming semantic clusters based on data stored on these peers, and then merging zones and clusters recursively until global zones and clusters are obtained.

3.1 Decentralized and Distributed Cluster Creation

The peer clustering process is divided into 5 phases: 1) local clustering, 2) zone initiator selection, 3) zone creation, 4) intra-zone clustering, and 5) inter-zone clustering.

Phase 1: Local Clustering. In the process of determining sites that contain related documents, *feature vectors* are used instead of the actual documents because of the large amounts of data involved. A feature vector F_i is a vector of tuples, each tuple containing a feature (word) f_i and a weight w_i. The feature vectors are created using a feature extraction process (more on the feature extraction process in section 6). By performing clustering of the document collection at each site, a set of document clusters is created, each cluster represented by a feature vector.

Phase 2: Initiator Selection. In order to be able to create zones, a subset of the peers have to be designated the role of *zone initiators* that can perform the zone creation process and subsequently initiate and control the clustering process within the zone.

The process of choosing initiators is completely distributed and ideally would be performed at all peers concurrently, in order to have approximately S_Z peers in each zone[1]. However, this concurrency is not necessary, since the use of zone partitioning at the next phase eliminates the danger of excessive zone sizes.

Assuming the IP of a peer P_i is IP_{P_i} and the time is T (rounded to nearest t_a[2]), a peer will discover that it is an initiator if $(IP_{P_i} + T)\ MOD\ S_Z = 0$. The aim of the function is to select initiators that are uniformly spread out in the network and an appropriate

[1] In order to avoid some initiators being overloaded, the aim is to have as uniform zone sizes as possible. Note that although uniform zone size and having initiator in the center of the zone are desired for load-balancing reasons, this is not crucial for the correctness or quality of the overlay construction.

[2] Assuming that each peer has a clock that is accurate within a certain amount of time t_a, note that DESENT itself can be used to improve the accuracy.

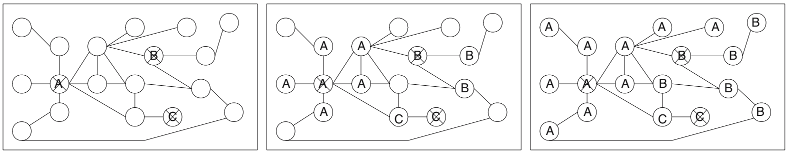


Fig. 1. Step-wise zone creation given the three initiators A, B, and C

number of initiators relative to the total number of peers in the network. By including time in the function we ensure that we obtain different initiators each time the clustering algorithm is run. This tackles the problem of being stuck with faulty initiators, as well as reduces the problem of permanent cheaters.

If no initiator is selected by the above strategy, this will be discovered from the fact that the subsequent zone creation phase is not started within a given time (i.e., no message received from an initiator). In this case, a universal decrease of the modulo-parameter is performed, by dividing by an appropriate prime number, as many times as necessary, in order to increase the chance of selecting (at least) one peer at the next iteration.

Phase 3: Zone Creation. After a peer P_i has discovered that it is an initiator, it uses a probe-based technique to create its zone. An example of zone creation is illustrated in Fig. 1. This zone creation algorithm has a low cost wrt. to number of messages (see Section 5), and in the case of excessive zone sizes, the initiator can decide to partition its zone, thus sharing its load with other peers. When this algorithm terminates, 1) each initiator has assembled a set of peers Z_i and their capabilities, in terms of resources they possess, 2) each peer knows the initiator responsible for its zone and 3) each initiator knows the identities of its neighboring initiators. An interesting characteristic of this algorithm is that it ensures that all peers in the network will be contacted, as long as they are connected to the network. This is essential, otherwise there may exist peers whose content will never be retrieved. We refer to the extended version of this paper for more details on initiator selection and zone creation [7].

Phase 4: Intra-zone Clustering. After the zones and their initiators have been determined, global clustering starts by collecting feature vectors from the peers (one feature vector for each cluster on a peer) and creating clusters based on these feature vectors:

1. The initiator of each zone i sends probe messages *FVecProbe* to all peers in Z_i.
2. When a peer P_i receives a *FVecProbe* it sends its set of feature vectors $\{F\}$ to the initiator of the zone.
3. The initiator performs clustering on the received feature vectors. The result is a set of clusters represented by a new set of feature vectors $\{F_i\}$, where an F_i consists of the top-k features of cluster C_i. Note that a peer can belong to more than one cluster. In order to limit the computations that have to be performed in later stages at other peers, when clusters from more than one peer have to be considered, the clustering should result in at most N_C^0 such basic clusters (N_C^0 is controlled by the clustering algorithm). The result of this process is illustrated in the left part of Fig. 2.

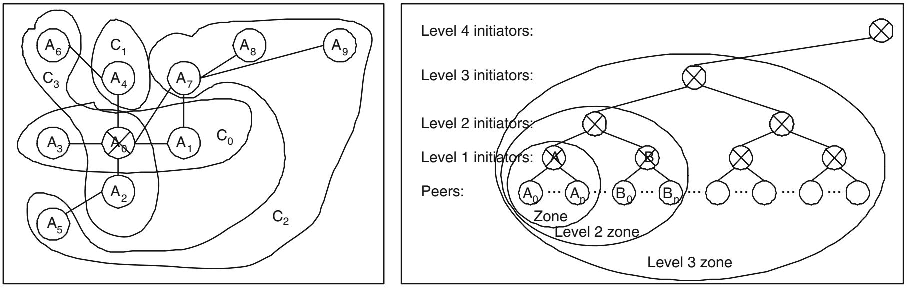


Fig. 2. Left: Possible result of intra-zone clustering of zone A, resulting in the four clusters C_0, C_1, C_2, and C_3. Right: Hierarchy of zones and initiators

4. The initiator selects a representative peer R_i for each cluster, based on resource information that is provided during Phase 3, like peer bandwidth, connectivity, etc. One of the purposes of a representative peer is to represent a cluster at search time.
5. The result kept at the initiator is a set of cluster descriptions (CDs), one for each cluster C_i. A CD consists of the cluster identifier C_i, a feature vector F_i, the set of peers $\{P\}$ belonging to the cluster, and the representative R of the cluster, i.e., $\text{CD}_i = (C_i, F_i, \{P\}, R)$. For example, the CD of cluster C_2 in Fig. 2 (assuming A_7 is the cluster representative) would be $\text{CD}_2 = (C_2, F_2, \{A_5, A_7, A_8, A_9\}, A_7)$.
6. Each of the representative peers are informed by the initiator about the assignment and receive a copy of the CDs (of *all* clusters in the zone). The representatives then inform peers on their cluster membership by sending them messages of the type (C_i, F_i, R).

Phase 5: Inter-zone Clustering. At this point, each initiator has identified the clusters in its zone. These clusters can be employed to reduce the cost and increase the quality of answers to queries involving the peers in one zone. However, in many cases peers in other zones will be able to provide more relevant responses to queries. Thus, we need to create an overlay that can help in routing queries to clusters in remote zones. In order to achieve this, we recursively apply merging of zones to larger and larger super-zones, and at the same time merge clusters that are sufficiently similar into super-clusters: first a set of neighboring zones are combined to a super-zone, then neighboring super-zones are combined to a larger super-zone, etc. The result is illustrated in the right part of Fig. 2 as a hierarchy of zones and initiators. Note that level-i initiators are a subset of the level-$(i-1)$ initiators.

This creation of the inter-zone cluster overlay is performed as follows:

1. From the previous level of zone creation, each initiator maintains knowledge about its neighboring zones (and their initiators). Thus, the zones essentially form a zone-to-zone network resembling the P2P network that was the starting point.
2. A level-i zone should consist of a number of neighboring level-$(i-1)$ zones, on average $|SZ|$ in each (where SZ denotes a set of zones, and $|SZ|$ the number of zones in the set). This implies that $\frac{1}{|SZ|}$ of the level-$(i-1)$ initiators should be level-i initiators. This is achieved by using the same technique for initiator selection

as described in Phase 2, except that in this case only peers already chosen to be initiators at level-$(i-1)$ in the previous phase are eligible for this role.

3. The level-i initiators create super-zones using the algorithm of Phase 3. In the same way, these level-i initiators will become aware of their neighboring super-zones.
4. In a similar way to how feature vectors were collected during the basic clustering, the approximately $N_C|SZ|$ CDs created at the previous level are collected by the level-i initiator (where N_C denotes the number of clusters per initiator at the previous level). Clustering is performed again and a set of super-clusters is generated. Each of the newly formed super-clusters is represented by top-k features produced by merging the top-k feature vectors of the individual clusters. The result of cluster merging is a set of super-clusters. A peer inside the super-cluster (not necessarily one of the representatives of the cluster) is chosen as representative for the super-cluster. The result is a new set of CDs, CD$_i$ $= (C_i, F_i, \{P\}, R)$, where the set of peers {P} contains the representatives of the clusters forming the base of the new super-cluster.
5. The CDs are communicated to the appropriate representatives. The representatives of the merged clusters (the peers in $\{P\}$ in the new CDs) are informed about the merging by the super-cluster representative, so that all cluster representatives know about both their representatives *below* as well as the representative *above* in the hierarchy. Note that although the same information could be obtained by traversing the initiator/super-initiator hierarchy, the use of cluster representatives distributes the load more evenly and facilitates efficient searching.

This algorithm terminates when only one initiator is left, i.e., when an initiator has no neighbors. Unlike the initiators at the previous levels that performed clustering operations, the only purpose of the final initiator is to decide the level of the final hierarchy. The aim is to have at the top level a number of initiators that is large enough to provide load-balancing and resilience to failures, but at the same time low enough to keep the cost of exchanging clustering information between them during the overlay creation to a manageable level. Note that there can be one or more levels below the top-level initiator that have too few peers. The top-level peer probes level-wise down the tree in order to find the number of peers at each level until it reaches level j with appropriate number min_F of peers. The level-j initiators are then informed about the decision and they are given the identifiers of the other initiators at that level, in order to send their CDs to them. Finally, all level-j initiators have knowledge about the clusters in zones covered by the other level-j initiators.

3.2 Final Organization

To summarize, the result of the zone- and cluster-creation process are two hierarchies:

Hierarchy of peers: Starting with individual peers at the bottom level, forming zones around the initiating peer which acts as a zone controller. Neighboring zones recursively form super-zones (see right part of Fig. 2), finally ending up in a level where the top of the hierarchies have replicated the cluster information of the other initiators at that level. This is a forest of trees. The peers maintain the following information about the rest of the overlay network: 1) Each peer knows its initiator. 2) A level-1 initiator knows the peers in its zone as well as the level-2 initiator of the super-zone it is covered by. 3) A level-i initiator (for $i > 1$) knows the identifiers of the level-$(i-1)$ initiators of the zones that constitute the super-zone as well as the level-$(i+1)$ initiator of the super-zone it is covered by. 4) Each initiator knows all cluster representatives in its zone.

Hierarchy of clusters: Each peer is member of one or more clusters at the bottom level. Each cluster has one of its peers as representative. One or more clusters constitute a super-cluster, which again recursively form new super-clusters. At the top level a number of global clusters exist. The peers store the following information about the cluster hierarchy: 1) Each peer knows the cluster(s) it is part of, and the representative peers of these clusters. 2) A representative also knows the identifiers of the peers in its cluster, as well as the identifier of the representative of the super cluster it belongs to. 3) A representative for a super-cluster knows the identifier of the representative at the level above as well as the representatives of the level below.

3.3 Peer Join

A peer P_J that joins the network first establishes connection to one or more peers as part of the basic P2P bootstrapping protocol. These neighbors provide P_J with their zone initiators. Through one of these zone initiators, P_J is able to reach one of the top-level nodes in the zone hierarchy and through a search downwards find the most appropriate lowest-level cluster, which P_J will then subsequently join. Note that no reclustering will be performed, so after a while a cluster description might not be accurate, but that cannot be enforced in any way in a large-scale, dynamic peer-to-peer system, given the lack of total knowledge. However, the global clustering process is performed at regular intervals and will then create a new clustering that reflects also the contents of new nodes (as well as new documents that have changed the individual peer's feature vectors). This strategy considerably reduces the maintenance cost, in terms of communication bandwidth compared with incremental reclustering, and also avoids the significant cost of continuous reclustering.

4 Searching

In this section we provide an overview of query processing in DESENT. A query Q in the network originates from one of the peers P, and it is continually expanded until satisfactory results, in terms of number and quality, have been generated. All results that are found as the query is forwarded are returned to P. Query processing can terminate at any of the steps below if the result is satisfactory:

1. The query is evaluated locally on the originating peer P.
2. A peer is a member of one or more clusters C_i. The C_i which has the highest similarity $sim(Q, C_i)$ with the query is chosen, and the query is sent to and evaluated by the other peers in this cluster.
3. Q is sent to one of the top-level initiators (remember that each of the top-level initiators knows about all the top-level clusters). At this point we employ two alternatives for searching:

Table 1. Parameters and default values used in the cost models

	Parameter	Default Value		Parameter	Default Value
B	Minimum bandwidth available	1 KB/s	N_i	# of peers/zones at level i	$\frac{N_P}{(S_Z)^i}$
D_0	Avg. # of neighbors at level 0	4	N_P	Total # of peers in the network	1000000
D_i	Avg. # of neighbors at level i	S_Z	r	Max zone radius	20
L	# of initiator levels	$\lfloor \log_{S_Z} N_P \rfloor$	S_{CD}	Size of a CD	$\approx 1.5 S_F$
min_F	Min. # of trees in top-level forest	$S_Z/4$	S_F	Size of feature vector	200 bytes
N_C^0	# of clusters per peer	10	S_M	Size of packet overhead	60 bytes
N_C^i	# of clusters per level-i initiator	100	S_Z	Avg. zone size	100
N_F	# of trees in top-level forest	$> S_Z/4$	t_a	Time between synch. points	60 seconds

(a) The most appropriate top-level cluster is determined based on a similarity measure, and Q is forwarded to the representative of that cluster. Next, Q is routed down the cluster hierarchy until the query is actually executed at the peers in a lowest-level cluster. The path is chosen based on highest $sim(Q, C_i)$ of the actual sub-clusters of a level-i cluster. If the number of results is insufficient, then backtracking is performed in order to extend the query to more clusters.

(b) All top-level clusters that have some similarity $sim(Q, C_i) > 0$ to the query Q are found and the query is forwarded to *all* cluster representatives. The query is routed down at *all* paths of the cluster hierarchy until level-0. Practically, all subtrees that belong to a matching top-level cluster are searched extensively.

The first approach reduces query latency, since the most relevant subset of peers will be identified with a small cost of messages. However, the number of returned documents will probably be restricted, since the search will focus on a local area only. This approach is more suitable for top-k queries. The second approach can access peers residing at remote areas (i.e. remote zones), with acceptable recall, however this results in a larger number messages. It is more suitable for cases when we are interested in the completeness of the search (retrieval of as many relevant documents as possible). In the following, we provide simulation results only for the second scenario, since we are mainly interested in testing the recall of our approach.

5 Feasibility Analysis

We have studied the feasibility of applying DESENT in a real-world P2P system through analytical cost models. Due to lack of space, we present here only the main results of the analytical study, whereas the actual cost models are described in detail in the extended version of this paper [7]. The parameters and default values used in the cost models are summarized in Table 1. These are typical values (practically size and performance) or values based on observations and conclusions from simulations.

A very important concern is the burden the DESENT creation imposes on participating nodes. We assume that the communication cost is the potential bottleneck and hence the most relevant metric, and we consider the cost of creating DESENT acceptable if the cost it imposes is relatively small compared to the ordinary document-delivery load on a web server.

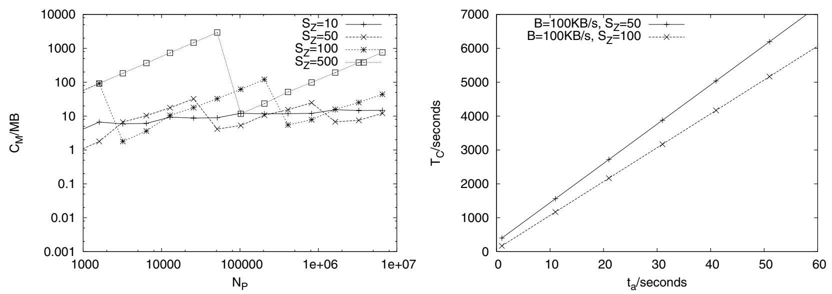


Fig. 3. Left: maximum cost of participation in overlay network creation for different values of network size N_P and zone size S_Z. Right: Time T_C to create DESENT as a function of t_a for different zone sizes and bandwidths.

In studying the feasibility of DESENT, it is important that the *average* communication cost for each peer is acceptable, but most important is the *maximum* cost that can be incurred for a peer, i.e., the cost for the initiators on the top level of the hierarchy. In order to study the maximum cost C_M for a particular peer to participate in the creation of the overlay network, both received and sent data should be counted because both pose a burden on the peer. Fig. 3 (left) illustrates C_M for different values of N_P and zone size S_Z. We see that a large zone size results in higher cost, but with very high variance. The situations in which this happens, is when the number of top-level peers is just below the min_F threshold so that the level below will be used as top level instead. With a large zone size this level will contain a large number of peers, and the final exchange of clusters information between the roots of this forest will be expensive. However, in practice this could be solved by merging of zones at this level. Regarding the maximum cost, if we consider a zone size of $S_Z = 100$, the maximum cost is just above 100 MB. Compared with the load of a typical web server, which is some GB of delivered documents per day, [3] this is acceptable even in the case of daily reclustering. However, considering the fact that the role of the upper-level initiators changes every time the overlay network is created, it could even be feasible to perform this clustering more often. In addition to the cost described above, there will also be a certain cost for maintaining replicas and peer dynamics in the network. However, this cost will be relatively small compared to the upper-level exchange of CDs.

In order to ensure freshness of the search results, it is important that the duration of the DESENT creation itself is not too long. The results, illustrated in Fig. 3 (right), show the time needed to create DESENT for different values of maximum assumed clock deviation, zone size S_Z, and minimum available bandwidth for DESENT participation B. For typical parameter values and $t_a = 30s$, the time needed to construct the DESENT overlay network is between 3000 and 4000 seconds, i.e., approximately one hour. This means that the DESENT creation could run several times a day, if desired. An important point is that even if the construction takes a certain time, the average load the construction imposes on peers will be relatively low. Most of the time is used to ensure

[3] Using a web server in our department as example, it delivers in the order of 4 GB per day, and a large fraction of this data is requested by search engines crawling the web.

that events are synchronized, without having to use communication for this purpose. Regarding values of parameters, it should be stressed that the actual *number of peers* has only minimal impact on the construction time, because the height of the tree is the important factor, and this increases only logarithmically with the number of peers.

6 DESENT Simulation Results

We have developed a simulation environment in Java, which covers all intermediate phases of the overlay network generation as well as the searching part. We ran all our experiments on Pentium IV computers with 3GHz processors and 1-2GB of RAM.

At initialization of the P2P network, a topology of N_P interconnected peers is created. We used the GT-ITM topology generator[4] to create random graphs of peers (we also used power-law topologies with the same results, due to the fact that the underlying topology only affects the zone creation phase), and our own SQUARE topology, which is similar to GT-ITM, only the connectivity degree is constant and neighboring peers share 3-5 common neighbors, i.e., the network is more dense than GT-ITM. A collection of N_D documents is distributed to peers, so that each peer retains N_D/N_P distinct documents. Every peer runs a clustering algorithm on its local documents resulting in a set of initial clusters. In our experiments we chose the Reuters-21578 text categorization test collection,[5] and we used 8000 pre-classified documents that belong to 60 distinct categories, as well as a different setup of 20000 documents. We tried different experimental setups with 2000, 8000 and 20000 peers. We then performed feature extraction (tokenization, stemming, stop-word removal and finally keeping the top-k features based on their TF/IDF[6] value and kept a feature vector of top-k features for each document as a compact document description). Thus, each document is represented by a top-k feature vector. Initiators retrieve the feature vectors of all peers within their zone, in order to execute intra-zone clustering. We used hierarchical agglomerative clustering (HAC) to create clusters of documents. Clustering is based on computing document similarities and merging feature vectors, by taking the union of the clusters' features and keeping the top-k features with higher TF/IDF values. We used the cosine similarity with parameter the similarity threshold T_s for merging. Clusters are created by grouping together sufficiently similar documents and each cluster is also represented by a top-k feature vector. Obviously, other clustering algorithms, as well as other similarity measures can be used.

6.1 Zone Creation

We studied the average zone size after the zone creation phase at level 1. The network topology consists of $N_P = 20000$ peers, each having 10 neighbors on average and

[4] `http://www.cc.gatech.edu/projects/gtitm/`

[5] `http://www.daviddlewis.com/resources/testcollections/reuters21578/`

[6] Notice that the inverse document frequency (IDF) is not available, since no peer has global knowledge of the document corpus, so we use the TF/IDF values produced on each peer locally, taking only the local documents into account.

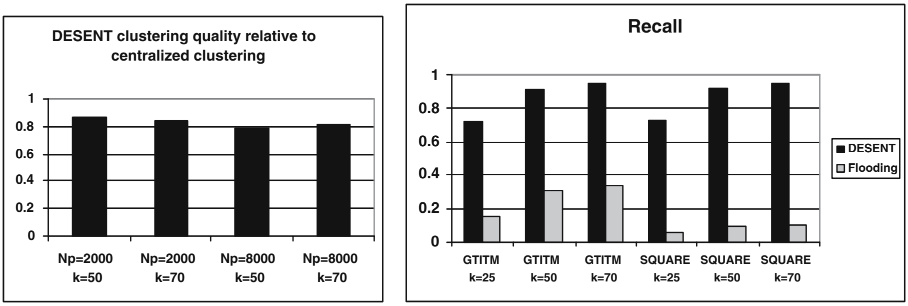


Fig. 4. Simulation results: Cluster quality, compared to centralized clustering, for different network sizes and values of k (left), and average recall compared to normalized flooding using the same number of messages (right)

$S_Z = 100$. We run the experiment with and without the zone partitioning mechanism. The simulations brought out the value of zone partitioning, since this mechanism keeps all zones smaller than S_Z, while most are of sizes $50 - 100$. However, when there is no zone partitioning, about 30% of the total zones have sizes greater than S_Z, and some are twice larger than S_Z, thus imposing a cumbersome load on several initiators.

6.2 Clustering Results Quality

Measuring the quality of the DESENT clustering results is essential for the value of the approach. As clustering quality in our context, we define the similarity of the results of our clustering algorithm (C_i), with respect to an optimal clustering (K_j). We used in our experiments the F-measure as a cluster quality measure. F-measure ranges between 0 and 1, with higher values corresponding to better clustering.

We compare the clustering quality of our approach to the centralized clustering results. The average values of DESENT F-measure relative to centralized clustering are illustrated in the left part of Fig 4, and show that DESENT achieves high clustering quality. Also note that the results exhibit a relatively stable behavior as the network size increases. This indicates that DESENT scales well with the number of participating peers. This conveys that the proposed system achieves high quality in forming SONs despite of the lack of global knowledge and the high distribution of the content.

6.3 Quality and Cost of Searching

In order to study the quality of searching in DESENT, we consider as baseline the search that retrieves all documents that contain all keywords in a query. We measure the searching quality using recall, representing the percentage of the relevant documents found. Note that, for the assumed baseline, precision will always be 100% in our approach, since the returned documents will always be relevant, due to the exact matching of all keywords. We generated a synthetic query workload consisting of queries with term count average 2.0 and standard deviation 1.0. We selected query terms from the documents randomly (ignoring terms with frequency less than 1%). The querying peer was selected randomly.

In the right part of Fig. 4, we show the average recall of our approach compared to normalized flooding using the same number of messages for different values of k, for the GT-ITM topology and the SQUARE topology for 8000 peers. Normalized flooding [8] is a variation of naive flooding that is widely used in practice, in which each peer forwards a query to d neighbors, instead of all neighbors, where d is usually the minimum connectivity degree of any peer in the network. The chart shows that with the same number of messages, our approach improves recall by more than 3-5 times for GT-ITM, and more than 10 for SQUARE, compared to normalized flooding. Furthermore, the absolute recall values increase with k, since more queries can match the enriched (with more features) cluster descriptions. Also notice that our approach presents the same recall independent of the underlying network topology.

7 Conclusions and Further Work

In this paper, we have presented algorithms for distributed and decentralized construction of hierarchical SONs, for supporting searches in a P2P-based digital library context. Future work includes performance and quality measurement of the search algorithm using large document collections, studying the use of other clustering algorithms as well as the use of caching techniques and ranking to increase efficiency.

Acknowledgments. The authors would like to thank George Tsatsaronis and Semi Koen for their help in preparing the feature extraction and clustering modules.

References

1. W.-T. Balke, W. Nejdl, W. Siberski, and U. Thaden. DL meets P2P - Distributed Document Retrieval based on Classification and Content. In *Proceedings of ECDL'2005*, 2005.
2. Y. Chawathe, S. Ratnasamy, L. Breslau, N. Lanham, and S. Shenker. Making Gnutella-like P2P Systems Scalable. In *Proceedings of SIGCOMM'03*, 2003.
3. E. Cohen, H. Kaplan, and A. Fiat. Associative Search in Peer-to-Peer Networks: Harnessing Latent Semantics. In *Proceedings of INFOCOM'03*, 2003.
4. A. Crespo and H. Garcia-Molina. Routing Indices for Peer-to-Peer Systems. In *Proceedings of ICDCS'2002*, 2002.
5. A. Crespo and H. Garcia-Molina. Semantic Overlay Networks for P2P Systems. Technical report, Stanford University, 2002.
6. H. Ding and I. Sølvberg. Choosing Appropriate Peer-to-Peer Infrastructure for your Digital Libraries. In *Proceedings of ICADL'2005*, 2005.
7. C. Doulkeridis, K. Nørvåg, and M. Vazirgiannis. DESENT: Decentralized and Distributed Semantic Overlay Generation in P2P Networks. Technical report, AUEB, 2005 `http://www.db-net.aueb.gr/index.php/publications/technical_reports/`.
8. C. Gkantsidis, M. Mihail, and A. Saberi. Hybrid Search Schemes for Unstructured Peer-to-Peer Networks. In *Proceedings of INFOCOM'05*, 2005.
9. Y. Ioannidis, H.-J. Schek, and G. Weikum, editors. *Proceedings of the 8th International Workshop of the DELOS Network of Excellence on Digital Libraries on Future Digital Library Management Systems (System Architecture & Information Access)*, 2005.
10. X. Liu, J. Wang, and S. T. Vuong. A Category Overlay Infrastructure for Peer-to-Peer Content Search. In *Proceedings of IPDPS'05*, 2005.

11. H. Nottelmann and N. Fuhr. Comparing Different Architectures for Query Routing in Peer-to-Peer Networks. In *Proceedings of ECIR'2006*, 2006.
12. L. Ramaswamy, B. Gedik, and L. Liu. Connectivity based Node Clustering in Decentralized Peer-to-Peer Networks. In *Proceedings of P2P'03*, 2003.
13. H. T. Shen, Y. Shu, and B. Yu. Efficient Semantic-based Content Search in P2P Network. *IEEE Transactions on Knowledge and Data Engineering*, 16(7):813–826, 2004.
14. P. Triantafillou, C. Xiruhaki, M. Koubarakis, and N. Ntarmos. Towards High Performance Peer-to-Peer Content and Resource Sharing Systems. In *Proceedings of CIDR'03*, 2003.
15. B. Yang and H. Garcia-Molina. Designing a Super-Peer Network. In *Proceedings of ICDE'03*, 2003.

Reevaluating Access and Preservation Through Secondary Repositories: Needs, Promises, and Challenges

Dean Rehberger, Michael Fegan, and Mark Kornbluh

310 Auditorium MATRIX Michigan State University East Lansing, MI 48824-1120 USA
rehberge@msu.edu, mfegan@msu.edu, kornbluh@msu.edu

Abstract. Digital access and preservation questions for cultural heritage institutions have focused primarily on primary repositories — that is, around collections of discrete digital objects and associated metadata. Much of the promise of the information age, however, lies in the ability to reuse, repurpose, combine and build complex digital objects[1-3]. Repositories need both to preserve and make accessible primary digital objects, and facilitate their use in a myriad of ways. Following the lead of other annotation projects, we argue for the development of secondary repositories where users can compose structured collections of complex digital objects. These complex digital objects point back to the primary digital objects from which they are produced (usually with URIs) and augment these pointers with user-generated annotations and metadata. This paper examines how this layered approach to user generated metadata can enable research communities to move forward into more complex questions surrounding digital archiving and preservation, addressing not only the fundamental challenges of preserving individual digital objects long term, but also the access and usability challenges faced by key stakeholders in primary digital repository collections—scholars, educators, and students. Specifically, this project will examine the role that secondary repositories can play in the preservation and access of digital historical and cultural heritage materials with particular emphasis on streaming media.

1 Introduction

To date, digital access and preservation questions for cultural heritage institutions have focused chiefly on primary repositories — that is, around collections of discrete digital objects and associated metadata. From the Library of Congress' American Memory to the digital image collection of the New York Public Library to the University of Heidelberg Digital Archive for Chinese Studies to the digital collections at the National Library of Australia, millions of objects are being made available to the general public that were once only the province of the highly trained researcher. Students have unprecedented access to illuminated manuscripts, primary and secondary documents, art, sheet music, photographs, architectural drawings, ethnographic case studies, historical voices, video, and a host of other rich and varied resources. The rapid growth of primary materials available online is well documented, as are the challenges posed by the "deep web."

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 39 – 50, 2006.

Access is at issue as well as preservation. While access to the "deep web" resources is difficult for most internet users, access to items in even the most well established repositories is largely limited to search, browse, and view. Much of the promise of the information age, however, lies in the ability to reuse, repurpose, combine and build complex digital objects[1-3]. Repositories need both to preserve and make accessible primary digital objects, and facilitate their use in a myriad of ways. Following the EU-NSF DL all projects meeting in March 2002 in Rome, Dagobert Soergel outlined a framework for the development of digital libraries by proposing that DLs need to move beyond the "paper-based metaphors" that privilege the finding and viewing of documents to support new ways of doing intellectual work [4]. The framework offers, among others, seven key points for this transformation of digital libraries: one, DLs need to support collaboration and communities of users with tools; two, the tools must be able to process and present the materials in ways that "serve the user's ultimate purpose"; three, users need to build their own individual or community "information spaces through the process of selection, annotation, contribution, and collaboration"; four, the tools need to be easy to use and should automate as many processes as possible; five, users need to be able to retrieve complex objects and interrelated structures; six, developers need to do careful analysis of user tasks and needs; and finally, seven, key to this framework is also the need to support user training and education to enhance further exploration and use of digital libraries.

While this framework appears ambitious (and expensive), we propose the development of secondary repositories where users can compose structured collections of complex digital objects with easy to use tools. These complex digital objects point back to the primary digital objects from which they are produced (usually with URIs) and users can augment these pointers with user-generated annotations and metadata. In so doing, users can organize the objects they find from a variety of DLs, personalizing and contextualizing the objects. They can gather a variety of media and format types, providing a meaningful presentation for themselves and their communities of users as well as a portal back to the digital libraries to encourage further investigation and discovery. The key element to the tool set is to provide affordances that encourage users to improve their ability to access digital libraries and develop ontologies that make sense to their community[s] of users. Since information in a secondary repository is generated and layered outside of the controlling system of the primary repository, such contextualized metadata currently would not be proposed as a replacement for current practices and initiatives but as an enhancement that seeks to support the current paradigm shift in research from object to use, presentation to interaction.

This paper examines how this layered approach to user generated metadata can enable research communities to move forward into more complex questions surrounding digital archiving and preservation, addressing not only the fundamental challenges of preserving individual digital objects long term, but also the access and usability challenges faced by key stakeholders in primary digital repository collections—scholars, educators, and students. Specifically, this project will examine the role that secondary repositories can play in the preservation and access of digital historical and cultural heritage materials with particular emphasis on streaming media.

2 Paradigm Shift

> "Many of the digital resources we are creating today will be repurposed and re-used for reasons we cannot imagine today. . . . Digital technologies are shaping creation, management, preservation, and access in ways which are so profound that traditional methods no longer are effective. These changes will require a paradigm shift in research if it is to provide the innovations—whether theoretical, methodological or technical—necessary to underpin long term access to digital resources."[1]

Many researchers and scholars within the digital library community recognize that new and innovative research directions are required to stimulate research on the long-term management and preservation of digital media.[2] The reasons for the call for a paradigm shift in the Digital Library community's research agenda are simple and direct. While access to online resources has steadily improved in the last decade, online archives and digital libraries still remain difficult to use, particularly for students and novice users [5]. In some cases, large amounts of resources have been put into massive digitization initiatives that have opened rich archives of historical and cultural materials to a wide range of users. Yet the traditional cataloging and dissemination practices of libraries and archives make it difficult for these users to locate and use effectively these sources, especially within scholarly and educational contexts [6]. Many digital libraries around the country, large and small, have made admirable efforts toward creating user portals and galleries to enhance the usability of their holdings, but these results are often expensive and labor intensive, often speaking only directly to a small segment of users or giving limited options for user interactivity. Most popular is the user-generated collection (e.g., Main Memory Network, users create their image galleries [7]). While an important step forward, these initiatives often develop tools that can be used only within a single archive that developed the tool.

To address these problems and to initiate the paradigm shift, researchers have questioned the gulf that separates issues of access from issues of preservation. Preservation and access are no longer entirely thought of in terms of stand alone files or individual digital objects, but in terms of active use—how users find, use, reuse, repurpose, combine and build complex digital objects out of the objects they collect. This assumption relies on a more complex meaning for the term "access." Many scholars in the field have called for a definition of access that goes beyond search interfaces to the ability of users to retrieve information "in some form in which it can be read, viewed, or otherwise employed constructively"[6, 8, 9]. Access thus implies four related conditions that go beyond the ability to link to a network: 1) equity—the ability of "every citizen" and not simply technical specialists to use the resources; 2) usability—the ability of users to easily locate, retrieve, use, and navigate resources; 3) context—the conveyance of meaning from stored information to users, so that it makes sense to them; and 4) interactivity—the capacity for users to be both consumers and producers of information.

Researchers have noted that the keys to enhancing access for specific user groups, contexts, and disciplines are to build repositories with resources and tools that allow users to enhance and augment materials[10], share their work with a community of users[11], and easily manipulate the media with simple and intuitive tools. Users will also need portal spaces that escape the genre of links indexes and become flexible work environments that allow users to become interactive producers[12].

2.1 The Challenges of Metadata

Over the past decade, the digital library community has tried to reduce the labor and expense of creating, cataloging, storing, and disseminating digital objects through the research and development of specific practices to facilitate each of these stages. In the face of ever-accelerating rates of complex data-creation and primary repository development, the central challenge to the digital library community is the long term sustainability and cost-effectiveness of primary digital repositories. The greatest cost factor in the field of digital preservation is human labor, "with current methods relying on significant human intervention for selection, organization, description and access" [1]. Leaders in the field of digital preservation are asking how metadata, semantics, and knowledge management technologies can enable the future reuse of primary repository collections; while at the same time minimize the labor intensiveness of the process [2]. Although current processes have become easier, better documented, and more automated, creating and working with digital objects is still a very specialized endeavor that requires specialized hardware, software, and expertise. This expertise is for the most part outside of the realm and resources for many cultural institutions and small digital libraries.

In line with digital library best practices, digitized sources are typically cataloged to describe their bibliographic information, along with technical, administrative, and rights metadata. While these practices are essential for preserving the digital object and making it available to users, unfortunately they do so in a language and guise often difficult to understand within the context of use [3, 13]. As Hope Olson points out, traditional cataloguing practices based on LCSH and DDC, while essential to giving access to items, often disproportionately affects access for marginalized groups and topics falling outside of mainstream culture [14]**.** Similarly, even though the author's name, the title of the work, and keywords are essential for describing and locating a digital object, this kind of information is not always the most utilized information for ascertaining the relevance of a digital object. For instance, K-12 teachers often do not have specific authors or titles in mind when searching for materials for their classes. Teachers more frequently search in terms of grade level, the state and national standards that form the basis of their teaching, or broad overarching topics derived from the required content and benchmark standards (e.g., core democratic values or textbook topics) that tend to display too many search returns to make the information of value.

This problem for educators has been one of the primary reasons for the development of Learning Object Metadata (LOM) [15]. Through improved metadata attached to learning objects, the hope is that educators can more easily find, assemble, and use units of educational content. Using object-oriented programming as a metaphor, the emphasis is on avoiding needless replication of labor by assembling learning objects found on the internet to build course material. This approach has provided excellent resources, particularly for the sciences, math and engineering. Yet Paul

Shabajee has chronicled well the problems associated with learning object metadata [10]. While it can do an excellent job of facilitating access to learning objects, especially for well-developed models and simulations, for raw assets (images, video segments, audio clips) assigning learning object metadata can exclude as much as give access. For examples, a set of images of a New Hampshire village may be designed for a college-level course on ethnography, but could be used on any level for a number of subjects from art to history to social studies to architecture (an infinite variety of uses). Moreover, learning object repositories usually are either a collection of objects with no relation to other digital libraries (from which facets of the object may have been taken) or as a collection of link reviews. While instructors can assemble good materials for their classes, the materials are often in the form of sets of links that do not articulate or contextualize access to related digital libraries nor do they allow for much personalization or change.

Researchers have long grappled with the problems of costs, knowledge, and resources needed to do full cataloguing of digital objects. As is well known, the Dublin Core initiative directly addresses the problem by specifying a minimal set of metadata to enhance searching and indexing of digital objects. The Dublin Core has worked so well that studies are now demonstrating that authors can apply metadata to their creations as well as professional [16]. Similarly, taking advantage of the XML namespace, the Resource Description Framework provides a modular approach to metadata, allowing for the accommodation of numerous and varied metadata packages from a variety of user groups. While viable instantiations of RDF have been limited to specialized areas and commerce, it does provide a wrapper that would work well to exchange metadata between secondary repositories. Dublin Core (which could be harvested or submitted from participating digital repositories), provides for the initial metadata needed to create secondary repositories, their access and development, which is then enhanced by user-generated metadata.

2.2 The Challenges of Annotating Streaming Media

Even though access by specialist scholars and educators to digital objects has grown at an exponential rate, tangible factors have prevented them from fully taking advantage of these resources in the classroom, where they could provide the conceptual and contextual knowledge of primary objects for their students. When educators do find the materials they need, using objects from various primary repositories to put together presentations and resources for their students and research can be challenging. Beyond merely creating lists of links to primary and secondary resources, assembling galleries of images, segmenting and annotating long audio and video files require far more technical expertise and time than can realistically be expected in the educational context. Additionally, even though scholars have a long history of researching archives and are comfortable sifting through records, locating items, and making annotations, comparisons, summaries, and quotations, these processes do not yet translate into online tools. Contemporary bibliographic tools have expanded to allow these users to catalogue and keep notes about media, but they do not allow users to mark specific passages and moments in multimedia, segment it, and return to specific places at a later time. Multimedia and digital repository collections thus remain underutilized in education and research because the tools to manipulate the various formats often “frustrate would be users” and take too much cognitive effort and time to learn[17].

While cursory studies have indicated these access issues, still very little is known about archival use or how these users express their information needs [18, 19]. For digital libraries to begin to fulfill their potential, much research is needed to understand better the processes by which primary repositories are accessed and how information needs are expressed. For example, research needs to address the ways in which teachers integrate content into their pedagogy so that bridges can be built from digital repositories to the educational process, bridges that greatly facilitate the ability of teachers and students to access specific information within the pedagogical process. Recent research strongly suggests that students need conceptual knowledge of information spaces that allow them to create mental models to do strategic and successful searches. As with any primary source, the materials in digital libraries do not literally “speak” for themselves and impart wisdom; they require interpretation and analysis [20]. Allowing communities of users to enhance metadata and actively use, reuse, repurpose, combine and build complex digital objects can help users to contextualize the information they find, draw from deeper resources within the digital library, and find more meaningful relationships between digital objects and their needs. Thinking in terms of a distributed model (similar to the open source software community) that allows users both easier access to materials and a greater range of search criteria and also provides opportunity for active engagement in the generation of metadata and complex digital objects, promises to help us rethink our most basic assumptions about user access and long-term preservation.

Researchers have long recognized the importance of user generated annotations and developing ontologies for differing user communities. Relevance feedback from users and interactive query expansion have been used to augment successfully metadata for document and image retrieval. The annotation and Semantic Web communities have made great strides in developing semi-automated annotation tools to enhance searching for a variety of media. Although many of the developed tools (SHOE Knowledge Annotator, MnM annotation tool, and WebKB) focus on HTML pages, the CREAting Metadata for the Semantic Web (CREAM) annotation framework promises to support manual and semi-automated annotation of both the shallow and deep web through the development of OntoAnnotate [21]. Other annotation projects tend to focus on particular fields, G-Portal (geography) and ATLAS (linguistics) and support a number of user groups within the field. Several of these annotation projects have worked remarkably well within distinct, highly trained user groups, but are more problematic when used by untrained, general users or in fields with less highly defined ontologies.

The secondary repository that we have built draws on the lessons learned annotation community. It is responsible for handling secondary metadata, extended materials and resources, interactive tools and application services. This information is cataloged, stored, and maintained in a repository outside of the primary repository that holds the digital object. The comments and observations generated by users in this context are usually highly specialized because such metadata is created from discipline-specific, scholarly perspectives (as an historian, social scientist, teacher, student, enthusiast, etc.) and for a specific purpose (research, publishing, teaching, etc.). Affordances are built in to help users identify themselves and their fields of interest. Even though the information generated by a secondary repository directly relates to digital objects in primary repositories, secondary repositories remain distinctly separate from the traditional repository. The information gathered in secondary repositories would rarely be used in the primary cataloging and maintenance of the object, and primary repositories would continue to be responsible for preservation, management, and long-term access but could be freed from creating time-consuming and expensive materials, resources, services, and extended metadata for particular user groups.

MATRIX: Center for Humane Arts, Letters and Social Sciences OnLine, at Michigan State University, for instance, has created a secondary repository using a server-side application called MediaMatrix [22]. This application is an online tool that allows users to easily find, segment, annotate and organize text, image, and streaming media found in traditional online repositories. MediaMatrix works within a web browser, using the browser's bookmark feature, a familiar tool for most users. When users find a digital object at a digital library or repository, they simply click the MediaMatrix bookmark and it searches through the page, finds the appropriate digital media, and loads it into an editor. Once this object is loaded, portions of the media can be isolated for closer and more detailed work—portions of an audio or video clip may be edited into annotated time-segments, images may be cropped then enlarged to highlight specific details. MediaMatrix provides tools so that these media can be placed in juxtaposition, for instance, two related images, a segment of audio alongside related images and audio, and so forth. Most importantly, textual annotations can be easily added to the media, and all this information is then submitted and stored on a personal portal page.

This portal page can be created by a scholar-educator who wishes to provide specific and contextualized resources for classroom use, and/or by a student creating a multimedia-rich essay for a class assignment. While these users have the immediate sense that they are working directly with primary objects, it is important to emphasize that primary repository objects are not actually being downloaded and manipulated. MediaMatrix does not store the digital object, rather, it stores a pointer to the digital object (URI) along with time or dimension offsets the user specified for the particular object and the user's annotation for that particular object. This use of URI pointing as opposed to downloading is especially significant because it removes the possibility that items may be edited and critiqued in contexts divorced from their original repositories, which hold the primary and crucial metadata for such objects. As long as primary repositories maintain persistent URIs for their holdings the pointer to the original digital object will always remain within the secondary repository, which acts as a portal to both the primary collection and contextualizing and interpretive information generated by individuals on items in those collections. This information is stored in a relational database along with valuable information about the individual, who supplies a profile regarding their scholarly/educational background, and provides information of the specific purposes for this work and the user-group (a class, for example) accessing the materials. The secondary repository can thus be searched and utilized in any number of ways.

3 Secondary Repositories and the Sustainability of Primary Repositories

At its most basic level, a secondary repository provides four levels of information concerning the use of digital objects housed in the primary repository: what is being used; what portions of those files are most utilized; who is using the digital objects; and, for what purpose are they using it. This information may be utilized in a number of different ways to support preservation and migration practices and the long-term sustainability of digital archives. Secondary repositories can instantly generate a list of the digital objects being used from any primary repository. This information could be used in determining digitization and preservation strategies as materials that are being utilized most by users might be pushed up the migration schedule and materials similar to those being most utilized might be digitized ahead of those materials that are least used. Because secondary repositories like MediaMatrix also allow users to segment digital objects by storing the time parameters of the sections they use, secondary repositories reveal what parts of digital objects users are most frequently accessing. This is not only helpful in determining segmentation strategies for all files and whether to further create specific semantic/intellectually meaningful segments for specific files, it removes the need for segmentation by the primary repository altogether. Repositories can store the time offsets (for audio and video files) or dimension markers (for images) to dynamically create segments of whole digital objects by feeding the offsets to the appropriate media player when the digital object is streamed or downloaded.

Of key importance to digital libraries is the issue of getting digital access and preservation on the agenda of key stakeholders such as universities and education systems. This agenda must be presented in terms that they will understand, and the ability to provide information about whom from these various communities is accessing particular digital objects from their holdings and for what purpose they are using them will be invaluable. The information contained in secondary repositories can assist stewards of primary repositories in building galleries and portals of digital objects that pertain to the needs of specific populations of users. This enables a more targeted approach to funding and project development. Whereas most primary repositories have educational sections, limitations in resources and labor often means that they can typically only offer a limited number of lesson plans that have relatively few digital objects (in relation to whole collections) from the primary holdings associated with them. Secondary repositories may give curators of primary repositories a better glimpse into how a specific user-base is using their holdings. Digital libraries can package materials especially suited for a specific demographic as well as instantly offer “additional suggestions” via a qualitative recommender system (for example, “Social Science, Grade 10-12 Teachers who accessed this image also viewed these resources”). Secondary repositories can even offer suggestions and links to similar digital objects housed at other primary repositories, therefore offering a truly federated resource. Secondary repositories can not only directly impact the sustainability of long term preservation projects, but also provides fruitful areas for further research and development on how recommender systems can be used effectively in these contexts, and how users interact with digital objects and personalize and repurpose information within specific contexts for specific purposes.

In creating new models for making digital preservation affordable and attractive to individuals, government agencies, universities, cultural institutions, and society at large, secondary repositories can perform vital roles. By enhancing and increasing meaningful access to primary repository holdings and by providing tools for quantifying and assessing that access within specific groups and educational context, secondary repositories can raise public awareness of digital preservation needs and also attract key stakeholders such as universities, libraries, and government agencies to invest in the continuance of digital preservation and access.

3.1 Secondary Repositories and Metadata

Secondary repositories may also provide a wealth of extensive metadata that pertain to the digital objects to which they point. While many would discount the usefulness of this metadata since it is primarily user-generated and does not follow cataloging standards, like dirty transcripts that contain various kinds and levels of errors, the annotations and notes generated by users could be used as additional criteria for keyword searching. This metadata would not replace traditional descriptions, keywords, and subject headings developed by catalogers, but rather it would be used in tandem with this metadata. As noted above, the real utility of this metadata is that it is generated from a very discipline/user specific vantage point and speaks to the language and conventions of that group. Traditional finding tools (keyword searches, thematic browsing, galleries, etc.) are problematic to many segments of users, stemming not only from the user's inability to formulate effective searches or lack of knowledge, but also the metadata that is searched and used to create these utilities. Because the metadata generated from secondary repositories is created by the same kind of user who will eventually search for specific digital objects, it often speaks directly to the methods and language they will use with search and browse utilities. User-specific metadata sets can be created using user profiles so that scholars have the ability to search the traditional catalogs, but also search through the annotations created by others within their field. Teachers will be able to search through the metadata created by others teaching the same grade level and subject matter. While traditional metadata approaches need to remain driven by best practices and community standards, secondary repositories provide a way to augment this metadata with a very personalized method of finding information.

This personalized and organic approach to metadata will help archivists of primary repositories identify what types of information future generations will need to use archival records, and help us to begin to answer the question "what information will people need to be able to continuously use records across time?" Secondary repositories can thus raise interesting questions as to the very function of metadata and what it means to preserve an object. The object itself represents "the tip of a very large iceberg; the tip is visible above the water only because there is a large mass of complex social relationships 'underneath' it—that generate, use and give meaning to, the digital documents." The object itself is more effectively thought of as a principle of organization for a complex nexus of interactions, events, and conversations that give meaning to a particular object. But, as Wendy Duff asks, how would archivists begin to represent the context of these records? What types of metadata are needed to document these relationships? There are many levels of metadata that need to be addressed to catalog properly the creation, nature, and life of a digital object [19].

Descriptive, copyright, source, digital provenance, and technical metadata work to ensure digital repositories can properly manage, find, migrate, and disseminate digital objects. In a sense they track the life of a digital object (where it came from, its different manifestations, changes in copyright, etc.) and ensure its access in the future (descriptive metadata function will work as a traditional finding aids, and technical metadata will provide information on how to render the object). While these hooks into digital objects can never be replaced by user-generated metadata, other disciplines would argue that more is needed to preserve truly the life and meaning of an object over time. Indeed, social theorists would argue that a digital object's meaning is socially constructed through its use. Thus one way to begin to understand an object is to understand how people interpreted and used the object at a particular point in time. Similar to the marginalia written in books, interpretations of works of art or historical artifacts, translations of the now-lost dead sea scrolls, or the scribbled notes and diagrams in Watson and Cricke's workbooks, secondary repositories provide a unique way of documenting and preserving the meaning—and the construction of the meaning—of an object by revealing how specific users made meaning out the object at specific times and for specific uses. If the goal of preservation is to retain the truest sense of an object over time, this information would help define a richer sense of an object's meaning at any given time.

The preservation of metadata that works to preserve the meaning of a digital object over time is being broached in an indirect way through the development of Fedora (Flexible Extensible Digital Object and Repository Architecture - http://www.fedora.info/) and the use/development of METS (Metadata Encoding and Transmission Standard - http://www.loc.gov/standards/mets/). METS is a metadata standard that was specifically created to encode all levels of metadata needed to preserve, manage and disseminate a digital object. Fedora, which is an open-source digital object repository management system, uses METS as its primary metadata scheme. In its early conception, Fedora was struggling with using the METS scheme because it did not have a specific way of documenting the behaviors (definitions and mechanisms) FEDORA uses for each digital object. Behaviors are directions for doing something with a digital object and the parameters in order to perform that action. A sample behavior might be to get an image and display it at a specific size in a web browser. This information does not specifically describe the digital object; instead it provides instructions for computer applications on how to process the digital object in a particular way. The original inception of METS did not have an obvious place to store this information within the METS scheme. The creators of FEDORA successfully lobbied to have a section where multiple behaviors could be tied to a single digital object.

While this information is functional in the use and dissemination of digital objects, it also presents an interesting history of how specific digital objects were processed and presented to users over time. It documents the evolution of technology and how technology was used to present digital objects to users in a meaningful way. Secondary repositories would work much the same way by preserving which digital objects were selected and how users processed digital object in their own work. While repositories produce metadata that documents the nature and life of a digital object so that it can be managed and found, the difficult question remains: what other kinds of metadata are required so that multiple audiences can successfully use digital objects each in their own discipline-specific practices?

4 Conclusions

From this survey of work and our initial studies, we have found several serious research and community challenges still need to be broached.

Persistent URIs: URIs are an important aspect of secondary repositories and of tools for building secondary repositories like MediaMatrix. Digital archives and libraries have increasingly hindered access and re-access to digital objects by limiting access or granting temporary URIs for a digital object. While the importance of stable, persistent URIs has been well documented in the library community, they are especially important to secondary repositories. To respect the access restrictions built around digital objects by primary repositories, secondary repositories need to store a unique, persistent URI that allows the user to re-access the digital object they have annotated.

Standardizing Secondary Repository Metadata: While metadata standards have been thoroughly researched and developed for primary repositories, a standardized metadata scheme for secondary repositories has yet to be developed. Metadata, semantics, and knowledge management technologies need to enable future reuse of collections in digital archives. In particular, research and standardization are required for the metadata needed to help users make sense of objects, to help the secondary repository administrators manage the entries of users, and to preserve that information over time. The standardization of secondary repository metadata is especially important so metadata can be easily exchanged between secondary repository tools and between the secondary and primary repository. For secondary repositories to be truly useful for the user, they need to be able to use a number of different tools to work with and produce information about digital objects. To work with multiple tools developed by any number of institutions, a common approach to documenting and exchanging metadata needs to be adopted so that users can easily take their entries from one tool and import them into another. As noted above, this is one area in which substantial work has been done by the annotation community of researchers; porting and reevaluating of this work needs to be done in relation to cultural heritage materials and the humanities. It is also essential to produce a mutually beneficial relationship between secondary and primary repositories. Primary repositories need to access information easily that specifically pertains to the digital objects from their repository and utilize it within their own infrastructure. It will also be beneficial if secondary repositories can access and integrate small bits of metadata from the primary repository. This would provide the user with official metadata to accompany their annotations (helping to automate as much as the process as possible) as well as provide a means of detecting changes in URI's and updates to the digital object itself.

Preservation of complex digital objects: Primary repositories primarily change through the addition of digital objects to their holdings. The metadata for those digital objects is relatively static except for documenting the migration of those objects to different storage mediums or file formats. Secondary repositories on the other hand are organic, ever-changing entities. Users adjust the segments they have created, revise annotations and other information they have recorded for the object, and delete whole entries at will; they can restrict and allow very levels of access. Because these entries work to help preserve the meaning of a digital object over the time, questions arise as to how we will preserve the changing metadata created by users or whether we will preserve these changes at all. Like dirty transcripts, do we accept the flawed nature of the metadata created by users or are the changes made by users important bits of information in studying the use and evolution of the digital objects they describe?

Acknowledgements

Support for the project comes in large part due to the JISC/NSF Digital Libraries Initiative II: Digital Libraries in the Classroom Program—National Science Foundation, award no. IIS-0229808.

References

[1] M. Hedstrom, in Wave of the Future: NSF Post Digital Libraries Futures Workshop, Chatham, MA, 2003.

[2] N. S. F. a. t. L. o. Congress, in Workshop on Research Challenges in Digital Archiving and Long-Term Preservation (M. Hedstrom, ed.), 2003, p. 1.

[3] C. Lynch, in NSF Post Digital Libraries Futures Workshop, 2003.

[4] D. Soergel, in D-Lib Magazine, Vol. 8, 2002.

[5] W. Y. Arms, Digital libraries, MIT Press, Cambridge, Mass., 2000.

[6] C. L. Borgman, From Gutenberg to the global information infrastructure: access to information in the networked world, MIT Press, Cambridge, Mass., 2000.

[7] M. H. Society, Vol. 2004, 2001.

[8] C. Lynch, Wilson Library Bulletin 69 (1995) 38.

[9] B. Kahin, J. Keller, and Harvard Information Infrastructure Project., Public access to the Internet, MIT Press, Cambridge, Mass., 1995.

[10] P. Shabajee, in D-Lib Magazine, Vol. 8, 2000.

[11] R. Waller, in Ariadne, UKOLN, 2004.

[12] P. Miller, in Ariadne, 2001.

[13] C. Lynch, in Personalization and Recommender Systems in the Larger Context: New Directions and Research Questions (Keynote Speech). Dublin City University, Ireland, 2001.

[14] H. A. Olson, The power to name: locating the limits of subject representation in libraries, Kluwer Academic Publishers, Dordrecht, The Netherlands; Boston, 2002.

[15] T. I. L. T. S. C. (LTSC), Vol. 2003, IEEE, 2002.

[16] J. Greenberg, M. C. Pattuelli, B. Parsia, and W. D. Robertson, Journal of Digital Information 20 (2001).

[17] J. R. Cooperstock, in HCI International, Conference on Human-Computer Interaction, McGill, New Orleans, LA, 2001, p. 688.

[18] W. Duff and C. A. Johnson, American Archivist 64 (2001) 43.

[19] W. Duff, Archival Science (2001) 285.

[20] G. C. Bowker and S. L. Star, Sorting things out: classification and its consequences, MIT Press, Cambridge, Mass., 1999.

[21] S. Handschuh and S. Staab, in WWW2002, Honolulu, Hawaii, 2002.

[22] M. L. Kornbluh, D. Rehberger, and M. Fegan, in 8th European Conference, ECDL2004, Springer, Bath, UK, 2004, p. 329.

Repository Replication Using NNTP and SMTP

Joan A. Smith, Martin Klein, and Michael L. Nelson

Old Dominion University, Department of Computer Science
Norfolk, VA 23529 USA
{jsmit, mklein, mln}@cs.odu.edu

Abstract. We present the results of a feasibility study using *shared, existing*, network-accessible infrastructure for repository replication. We utilize the SMTP and NNTP protocols to replicate both the metadata and the content of a digital library, using OAI-PMH to facilitate management of the archival process. We investigate how dissemination of repository contents can be piggybacked on top of existing email and Usenet traffic. Long-term persistence of the replicated repository may be achieved thanks to current policies and procedures which ensure that email messages and news posts are retrievable for evidentiary and other legal purposes for many years after the creation date. While the preservation issues of migration and emulation are not addressed with this approach, it does provide a simple method of refreshing content with unknown partners for smaller digital repositories that do not have the administrative resources for more sophisticated solutions.

1 Introduction

We propose and evaluate two repository replication models that rely on *shared, existing* infrastructure. Our goal is not to “hijack” other sites’ storage, but to take advantage of protocols which have persisted through many generations and which are likely to be supported well into the future. The premise is that if archiving can be accomplished within a widely-used, already deployed infrastructure whose operational burden is shared among many partners, the resulting system will have only an incremental cost and be tolerant of dynamic participation. With this in mind, we examine the feasibility of repository replication using Usenet news (NNTP, [1]) and email (SMTP, [2]).

There are reasons to believe that both email and Usenet could function as persistent, if diffuse, archives. NNTP provides well-understood methods for content distribution and duplicate deletion (deduping) while supporting a distributed and dynamic membership. The long-term persistence of news messages is evident in “Google Groups,” a Usenet archive with posts dating from May 1981 to the present [3]. Even though blogs have supplanted Usenet in recent years, many communities still actively use moderated news groups for discussion and awareness. Although email is not usually publicly archivable, it is ubiquitous and frequent. Our departmental SMTP email server averaged over 16,000 daily

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 51–62, 2006.

outbound emails to more than 4000 unique recipient servers during a 30-day test period. Unlike Usenet, email is point-to-point communication but, given enough time, attaching repository contents to outbound emails may prove to be an effective way to disseminate contents to previously unknown locations. The open source products for news ("INN") and email ("sendmail" and "postfix") are widely installed, so including a preservation function would not impose a significant additional administrative burden.

These approaches do not address the more complex aspects of preservation such as format migration and emulation, but they do provide alternative methods for refreshing the repository contents to potentially unknown recipients. There may be quicker and more direct methods of synchronization for some repositories, but the proposed methods have the advantage of working with firewall-inhibited organizations and repositories without public, machine-readable interfaces. For example, many organizations have web servers which are accessible only through a VPN, yet email and news messages can freely travel between these servers and other sites without compromising the VPN. Piggybacking on mature software implementations of these other, widely deployed Internet protocols may prove to be an easy and potentially more sustainable approach to preservation.

2 Related Work

Digital preservation solutions often require sophisticated system administrator participation, dedicated archiving personnel, significant funding outlays, or some combination of these. Some approaches, for example Intermemory [4], Freenet [5], and Free Haven [6], require personal sacrifice for public good in the form of donated storage space. However, there is little incentive for users to incur such near-term costs for the long-term benefit of a larger, anonymous group. In contrast, LOCKSS [7] provides a collection of cooperative, deliberately slow-moving caches operated by participating libraries and publishers to provide an electronic "inter-library loan" for any participant that loses files. Because it is designed to service the publisher-library relationship, it assumes a level of at least initial out-of-band coordination between the parties involved. Its main technical disadvantage is that the protocol is not resilient to changing storage infrastructures. The rsync program [8] has been used to coordinate the contents of digital library mirrors such as the arXiv eprint server but it is based on file system semantics and cannot easily be abstracted to other storage systems. Peer-to-peer services have been studied as a basis for the creation of an archiving cooperative among digital repositories [9]. The concept is promising but their simulations indicated scalability is problematic for this model. The Usenet implementation [10] of the Eternity Service [11] is the closest to the methods we propose. However, the Eternity Service focuses on non-censorable anonymous publishing, not preservation per se.

3 The Prototype Environment

We began by creating and instrumenting a prototype system using popular, open source products: Fedora Core (Red Hat Linux) operating system; an NNTP news server (INN version 2.3.5); two SMTP email servers, postfix version 2.1.5 and sendmail version 8.13.1; and an Apache web server (version 2.0.49) with the mod_oai module installed [12]. *mod_oai* is an Apache module that provides Open Archives Protocol for Metadata Harvesting (OAI-PMH) [13] access to a web server. Unlike most OAI-PMH implementations, mod_oai does not just provide metadata about resources, it can encode the entire web resource itself in MPEG-21 Digital Item Declaration Language [14] and export it through OAI-PMH. We used Perl to write our own repository replication tools, which were operated from separate client machines.

As part of our experiment, we created a small repository of web resources consisting of 72 files in HTML, PDF and image (GIF, JPEG, and PNG) formats. The files were organized into a few subdirectories with file sizes ranging from less than a kilobyte to 1.5 megabytes. For the NNTP part of the experiment, we configured the INN news server with common default parameters: messages could be text or binary; maximum message life was 14 days; and direct news posting was allowed. For email, we did not impose restrictions on the size of outgoing attachments and messages. For each archiving method, we harvested the entire repository over 100 times.

Both the NNTP and SMTP methods used a simple, iterative process: (1)read a repository record; (2)format it for the appropriate archive target (mail or news); (3)encode record content using base64; (4)add human-readable X-headers (for improved readability and recovery); (5)transmit message (email or news post) to the appropriate server; (6)repeat steps 1 through 5 until the entire repository has been archived. Below, we discuss details of the differences in each of these steps as applied specifically to archiving via news or email.

We took advantage of OAI-PMH and the flexibility of email and news to embed the URL of each record as an X-Header within each message. X-Headers are searchable and human-readable, so their contents give a clue to the reader about the purpose and origin of the message. Since we encoded the resource itself in base 64, this small detail can be helpful in a forensic context. If the URL still exists, then the X-Headers could be used to re-discover the original resource. Table 1 shows the actual X-Headers added to each archival message.

3.1 The News Prototype

For our experiment, we created a *moderated* newsgroup which means that postings must be authorized by the newsgroup owner. This is one way newsgroups keep spam from proliferating on the news servers. We also restricted posts to selected IP addresses and users, further reducing the "spam window." For the experiment, we named our newsgroup "repository.odu.test1," but groups can have any naming scheme that makes sense to the members. For example, a DNS-based

Table 1. Example of Human-Readable X-Headers Added to Archival Messages

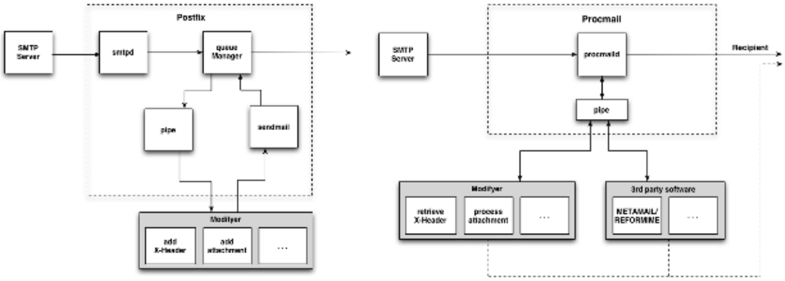
```
X-Harvest_Time: 2006-2-15T18:34:51Z
X-baseURL: http://beatitude.cs.odu.edu:8080/modoai/
X-OAI-PMH_verb: GetRecord
X-OAI-PMH_metadataPrefix: oai_didl
X-OAI-PMH_Identifier: http://beatitude.cs.odu.edu:8080/1000/pg1000-1.pdf
X-sourceURL: http://beatitude.cs.odu.edu:8080/modoai/?verb=GetRecord
&identifier=http://beatitude.cs.odu.edu:8080/1000/pg1000-1.pdf
&metadataPrefix=oai_didl
X-HTTP-Header: HTTP/1.1 200 OK
```

scheme that used “repository.edu.cornell.cs” or “repository.uk.ac.soton.psy” would be a reasonable naming convention.

Using the simple 6-step method outlined above, we created a news message with X-Headers for each record in the repository, We also collected statistics on (a)original record size vs. posted news message size; (b)time to harvest, convert and post a message; and (c)the impact of line length limits in news posts. Our experiment showed high reliability for archiving using NNTP. 100% of the records arrived intact on the target news server, “beatitude.” In addition, 100% of the records were almost instantaneously mirrored on a subscribing news server (“beaufort”). A network outage during one of the experiments temporarily prevented communication between the two news servers, but the records were replicated as soon as connectivity was restored.

3.2 The Email Prototype

The two sides of SMTP-method archiving, outbound and inbound, are shown in Figure 1. Archiving records by piggybacking on existing email traffic requires sufficient volume to support the effort and to determine which hosts are the best recipients. Analysis of outbound email traffic from our department during a 30-day period showed 505,987 outgoing messages to 4,081 uniquehosts. A power



(a) Outbound Mail (b) Inbound Mail

Fig. 1. Archiving Using SMTP

law relationship is also evident (see Figure 2) between the domain's rank and email volume sent to that domain:

$$V_\kappa = c * (\kappa^{-1.6}) \tag{1}$$

Using the Euler Zeta function (discussed in detail in [15]), we derived the value of the constant, $c = 7378$, in Equation 1.

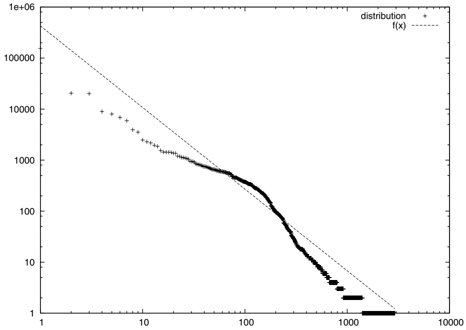


Fig. 2. Email distribution follows a power law

3.3 Prototype Results

Having created tools for harvesting the records from our sample digital library, and having used them to archive the repository, we were able to measure the results. How fast is each prototype and what penalties are incurred? In our email experiment, we measured approximately a 1 second delay in processing attachments of sizes up to 5MB. With NNTP, we tested postings in a variety of sizes and found processing time ranged from 0.5 seconds (12 KB) to 26.4 seconds (4.9MB). Besides the trivial linear relationship between repository size and replication time, we found that even very detailed X-Headers do not add a significant burden to the process. Not only are they small (a few bytes) relative to record size, but they are quickly generated (less than 0.001 seconds per record) and incorporated into the archival message. Both NNTP and SMTP protocols are robust, with most products (like INN or sendmail) automatically handling occasional network outages or temporary unavailability of the destination host. News and email messages are readily recovered using any of a number of "readers" (e.g., Pine for email or Thunderbird for news). Our experimental results formed the basis of a series of simulations using email and Usenet to replicate a digital library.

4 Simulating the Archiving Process

When transitioning from live, instrumented systems to simulations, there are a number of variables that must be taken into consideration in order to arrive at realistic figures (Table 2). Repositories vary greatly in size, rate of updates and additions, and number of records. Regardless of the archiving method, a repository will have specific policies ("Sender Policies") covering the number of copies archived; how often each copy is refreshed; whether intermediate updates are archived between full backups; and other institutional-specific requirements such as geographic location of archives and "sleep time" (delay) between the end of one completed archive task and the start of another. The receiving agent will have its own "Receiver Policies" such as limits on individual message size, length of time messages live on the server, and whether messages are processed by batch or individually at the time of arrival.

Table 2. Simulation Variables

Repository	R	Number of records in repository
	$R_{\overline{s}}$	Mean size of records
	R_a	Number of records added per day
	R_u	Number of records updated per day
	ρ	Number of records posted per day
Usenet	N_{ttl}	News post time-to-live
	S	"Sleep" time between baseline harvests
	ρ_{news}	Records postable per day via news
	T_{news}	Time to complete baseline using news
Email	G	Granularity
	κ	Rank of receiving domain
	c	Constant derived from Euler Zeta function
	ρ_{email}	Records postable per day via email
	T_{email}	Time to complete baseline using email

A key difference between news-based and email-based archiving is the active-vs-passive nature of the two approaches. This difference is reflected in the policies and how they impact the archiving process under each method. A "baseline," refers to making a complete snapshot of a repository. A "cyclic baseline" is the process of repeating the snapshot over and over again ($S = 0$), which may result in the receiver storing more than one copy of the repository. Of course, most repositories are not static. Repeating baselines will capture new additions (R_a) and updates (R_u) with each new baseline. The process could also "sleep" between baselines ($S > 0$), sending only changed content. In short, the changing nature of the repository can be accounted for when defining its replication policies.

4.1 Archiving Using NNTP

Figure 3 illustrates the impact of policies on the news method of repository replication. A baseline, whether it is cyclic or one-time-only, should finish before the end of the news server message life (N_{ttl}), or a complete snapshot will not be achieved. The time to complete a baseline using news is obviously constrained by the size of the repository and the speed of the network. NNTP is an older protocol, with limits on line length and content. Converting binary content to base64 overcomes such restrictions but at the cost of increased file size (one-third) and replication time.

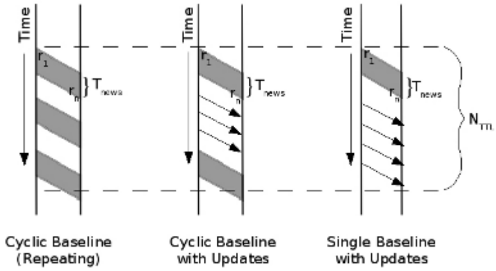

Time
Time
Time
r_1
r_n
T_{news}
r_1
r_n
T_{news}
r_1
r_n
T_{news}
N_{TTL}
Cyclic Baseline (Repeating)
Cyclic Baseline with Updates
Single Baseline with Updates

Fig. 3. NNTP Timeline for Sender & Receiver Policies

4.2 Archiving Using SMTP

One major difference in using email as the archiving target instead of news is that it is passive, not active: the email process relies on *existing* traffic between the archiving site and one or more target destination sites. The prototype is able to attach files automatically with just a small processing delay penalty. Processing options include selecting only every E^{th} email, a factor we call "granularity" [15]; randomly selecting records to process instead of a specific ordering; and/or maintaining replication lists for each destination site. Completing a baseline using email is subject to the same constraints as news - repository size, number of records, etc. - but is particularly sensitive to changes in email volume. For example, holidays are often used for administrative tasks since they are typically "slow" periods, but there is little email generated during holidays so repository replication would be slowed rather than accelerated. However, the large number of unique destination hosts means that email is well adapted to repository discovery through advertising.

5 Results

In addition to an instrumented prototype, we simulated a repository profile similar to some of the largest publicly harvestable OAI-PMH repositories. The simulation assumed a 100 gigabyte repository with 100,000 items ($R = 100000$, $R_{\overline{s}} = 1MB$); a low-end bandwidth of 1.5 megabits per second; an average daily update rate of 0.4% ($R_u = 400$); an average daily new-content rate of 0.1% ($R_a = 100$); and a news-server posting life (N_{ttl}) of 30 days. For simulating email replication, our estimates were based on the results of our email experiments: Granularity $G = 1$, 16866 emails per day, and the power-law factor applied to the ranks of receiving hosts. We ran the NNTP and SMTP simulations for the equivalent of 2000 days (5.5 years).

5.1 Policy Impact on NNTP-Based Archiving

News-based archiving is constrained primarily by the receiving news server and network capacity. If the lifetime of a posting (N_{ttl}) is shorter than the archiving time of the repository (T_{news}), then a repository cannot be successfully archived to that server. Figure 4 illustrates different repository archiving policies, where S ranges from 0 (cyclic baseline) to infinity (single baseline). The "Cyclic Baseline with Updates" in Figure 4 graphs a sender policy covering a 6-week period: The entire repository is archived twice, followed by updates only, then the cycle is repeated. This results in the news server having between one and 2 full copies of the repository, at least for the first few years. The third approach, where the policy is to make a single baseline copy and follow up with only updates and additions, results in a rapidly declining archive content over time, with only small updates existing on the server. It is obvious that as a repository grows and other factors such as news posting time remain constant, the archive eventually contains less than 100% of the library's content, even with a policy of continuous updates. Nonetheless, a significant portion of the repository remains archived for many years if some level of negotiated baseline archiving is established. As derived in [15], the probability of a given repository record r being currently replicated on a specific news server N on day D is:

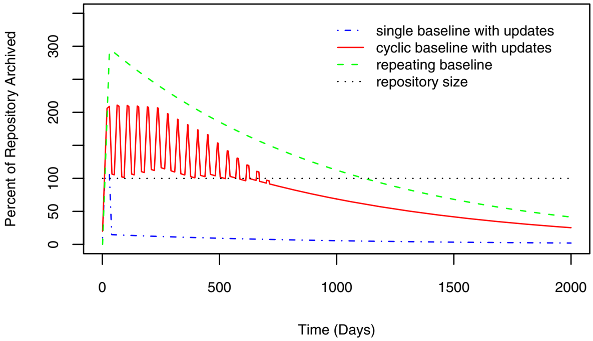


Fig. 4. Effect of Sender Policies on News-Method Archiving

$$P(r) = \frac{(\rho_{news} \times D) - \rho_{news} \times (D - N_{TTL})}{R + (D \times R_a)} \tag{2}$$

5.2 Policy Impact on SMTP-Based Archiving

SMTP-based replication is obviously constrained by the frequency of outbound emails. Consider the following two sender policies: The first policy maintains just one queue where items of the repository are being attached to every E^{th} email regardless of the receiver domain. In the second policy, we have more than one queue where we keep a pointer for every receiver domain and attach items to every E^{th} email going out to these particular domains. The second policy will allow the receiving domain to converge on 100% coverage much faster, since accidental duplicates will not be sent (which does happen with the first policy). However, this efficiency comes at the expense of the sending repository tracking separate queues for each receiving domain.

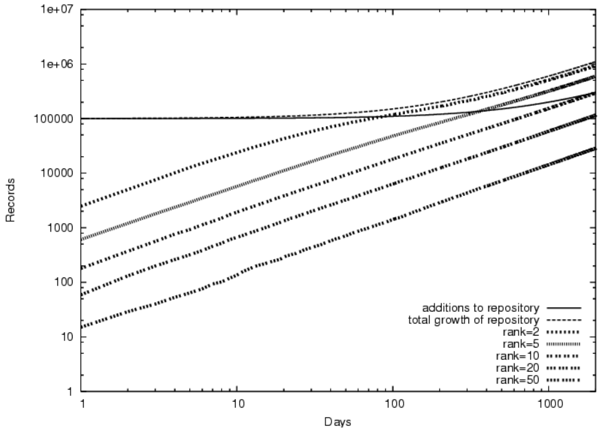
Because email volume follows a power law distribution, receiver domains ranked 2 and 3 achieve 100% repository coverage fairly soon but Rank 20 takes significantly longer (2000 days with a pointer), reaching only 60% if no pointer is maintained. Figure 5(a) shows the time it takes for a domain to receive all files of a repository without the pointer to the receiver and figure 5(b) shows the same setup but with receiver pointer. In both graphs, the 1^{st} ranked receiver domains are left out because they represent internal email traffic. Figure 5 shows how important record history is to achieving repository coverage using email. If a record history is not maintained, then the domain may receive duplicate records before a full baseline has been completed, since there is a decreasing statistical likelihood of a new record being selected from the remaining records as the process progresses. Thus, the number of records replicated per day via email ρ_{email} is a function of the receiver's rank (κ), the granularity (G), and probability based on use of a history pointer (h). That is, $\rho_{email} = c(\kappa^{-1.6}) * G * h$. If a pointer is maintained then $h = 1$; and if every outbound email to the domain is used, then $G = 1$ as well. The probability that a given record, r has been replicated via email is therefore:

$$P(r) = \frac{(\rho_{email} \times D)}{R + (D \times R_a)} \tag{3}$$

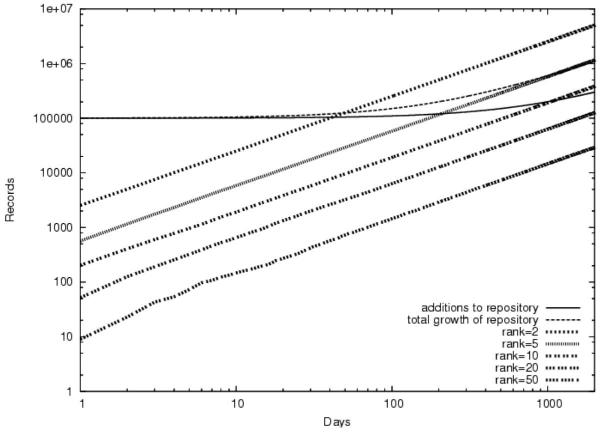
5.3 Discussion

How would these approaches work with other repository scenarios? If the archive were substantially smaller (10,000 records with a total size of 15 GB), the time to upload a complete baseline would also be proportionately smaller since replication time is linear with respect to the repository's size for both the news and email methods of archiving. The news approach actively iterates through the repository, creating its own news posts, and is therefore constrained primarily by bandwidth to the news server. Email, on the other hand, passively waits for existing email traffic and then "hitches a ride" to the destination host. The SMTP approach is dependent on the site's daily email traffic to the host, and a reduction in the number of records has a bigger impact if the repository uses the email solution because fewer emails will be needed to replicate the repository.

A repository consisting of a single record (e.g., an OAI-PMH "Identify" response) could be effectively used to advertise the existence of the repository regardless of the archiving approach or policies. After the repository was discovered, it could be harvested via normal means. A simple "Identify" record (in OAI-PMH terms) is very small (a few kilobytes) and would successfully publish the repository's existence in almost zero time regardless of the archiving approach that was used.



(a) Without Record History



(b) With Record History

Fig. 5. Time To Receive 100% Repository Coverage by Domain Rank

6 Future Work and Conclusions

Through prototypes and simulation, we have studied the feasibility of replicating repository contents using the installed NNTP and SMTP infrastructure. Our initial results are promising and suggest areas for future study. In particular, we must explore the trade-off between implementation simplicity and increased repository coverage. For SMTP approach, this could involve the receiving email domains informing the sender (via email) that they are receiving and processing attachments. This would allow the sender to adjust its policies to favor those sites. For NNTP, we would like to test varying the sending policies over time as well as dynamically altering the time between baseline harvests and transmission of update and additions. Furthermore, we plan to revisit the structure of the objects that are transmitted, including taking advantage of the evolving research in preparing complex digital objects for preservation [16][17].

It is unlikely that a single, superior method for digital preservation will emerge. Several concurrent, low-cost approaches are more likely to increase the chances of preserving content into the future. We believe the piggyback methods we have explored here can be either a simple approach to preservation, or a compliment to existing methods such as LOCKSS, especially for content unencumbered by restrictive intellectual property rights. Even if NNTP and SMTP are not used for resource transport, they can be effectively used for repository awareness. We have not explored what the receiving sites do with the content once it has been received. In most cases, it is presumably unpacked from its NNTP or SMTP representation and ingested into a local repository. On the other hand, sites with apparently infinite storage capacity such as Google Groups could function as long-term archives for the encoded repository contents.

Acknowledgements

This work was supported by NSF Grant ISS 0455997. B. Danette Allen contributed to the numerical analysis.

References

1. Brian Kantor and Phil Lapsley. Network news transfer protocol, Internet RFC-977, February 1986.
2. Jonathan B. Postel. Simple mail transfer protocol, Internet RFC-821, August 1982.
3. 20 year archive on google groups. `http://www.google.com/googlegroups/archive_announce_20.html`.
4. Andrew V. Goldberg and Peter N. Yianilos. Towards an archival intermemory. In *Proceedings of IEEE Advances in Digital Libraries, ADL 98*, pages 147–156, April 1998.
5. Ian Clark, Oskar Sandberg, Brandon Wiley, and Theodore W. Hong. Freenet: a distributed anonymous information storage and retrieval system. In *International Workshop on Design Issues in Anonymity and Unobservability LNCS 2009.*
6. Roger Dingledine, Michael J. Freedman, and David Molnar. The free haven project: Distributed anonymous storage service. *Lecture Notes in Computer Science*, 2009:67 –95, 2001.
7. Petros Maniatis, Mema Roussopoulos, T.J.Giuli, David S. H. Rosenthal, and Mary Baker. The LOCKSS peer-to-peer digital preservation system. *ACM Transactions on computer systems*, 23:2 – 50, February 2005.
8. Andrew Tridgell and Paul Mackerras. The rsync algorithm. Technical report, The Australian National University, 1996. `http://cs.anu.edu.au/techreports/1996/TR-CS-96-05.pdf`.
9. Brian F. Cooper and Hector Garcia-Molina. Peer-to-peer data trading to preserve information. *ACM Transactions on Information Systems*, 20(2):133 – 170, 2002.
10. Adam Back. The eternity service. *Phrack Magazine*, 7(51), 1997.
11. Ross J. Anderson. The eternity service. In *1st International Conference on the Theory and Applications of Cryptology (Pragocrypt '96)*, pages 242–252, 1996.

12. Michael L. Nelson, Herbert Van de Sompel, Xiaoming Liu, and Terry L. Harrison. mod_oai: An apache module for metadata harvesting. Technical report, Old Dominion University, 2005. arXiv cs.DL/0503069.
13. Carl Lagoze, Herbert Van de Sompel, Michael L. Nelson, and Simeon Warner. The Open Archives Initiative Protocol for Metadata Harvesting. `http://www.openarchives.org/OAI/openarchivesprotocol.html`.
14. Jeroen Bekaert, Patrick Hochstenbach, and Herbert Van de Sompel. Using MPEG-21 DIDL to represent complex digital objects in the Los Alamos National Laboratory digital library. *D-Lib Magazine*, 9(11), November 2003. doi:10.1045/november2003-bekaert.
15. Joan A. Smith, Martin Klein, and Michael L. Nelson. Repository replication using NNTP and SMTP. Technical report, Old Dominion University, 2006. arXiv cs.DL/0606008.
16. Jeroen Bekaert, Xiaoming Liu, and Herbert Van de Sompel. Representing digital assets for long-term preservation using MPEG-21 DID. In *Ensuring Long-term Preservation and Adding Value to Scientific and Technical data (PV 2005)*, 2005. arXiv cs.DL/0509084.
17. Herbert Van de Sompel, Michael L. Nelson, Carl Lagoze, and Simeon Warner. Resource harvesting within the OAI-PMH framework. *D-Lib Magazine*, 10(12), December 2004. doi:10.1045/december2004-vandesompel.

Genre Classification in Automated Ingest and Appraisal Metadata

Yunhyong Kim and Seamus Ross

Digital Curation Centre (DCC)
&
Humanities Adavanced Technology Information Institute (HATII)
University of Glasgow
Glasgow, UK

Abstract. Metadata creation is a crucial aspect of the ingest of digital materials into digital libraries. Metadata needed to document and manage digital materials are extensive and manual creation of them expensive. The Digital Curation Centre (DCC) has undertaken research to automate this process for some classes of digital material. We have segmented the problem and this paper discusses results in genre classification as a first step toward automating metadata extraction from documents. Here we propose a classification method built on looking at the documents from five directions; as an object exhibiting a specific visual format, as a linear layout of strings with characteristic grammar, as an object with stylo-metric signatures, as an object with intended meaning and purpose, and as an object linked to previously classified objects and other external sources. The results of some experiments in relation to the first two directions are described here; they are meant to be indicative of the promise underlying this multi-facetted approach.

1 Background and Objective

Construction of persistent, cost-contained, manageable and accessible digital collections depends on the automation of appraisal, selection, and ingest of digital material. Descriptive, administrative, and technical metadata play a key role in the management of digital collections ([37],[21]). As DELOS/NSF ([13],[14],[21]) and PREMIS working groups ([34]) noted metadata are expensive to create and maintain. Digital objects are not always accompanied by adequate metadata and the number of digital objects being created and the variety of such objects is increasing at an exponential rate. In response, the manual collection of metadata can not keep pace with the number of digital objects that need to be documented. It seems reasonable to conclude that automatic extraction of metadata would be an invaluable step in the automation of appraisal, selection, and ingest of digital material. ERPANET's ([17]) Packaged Object Ingest Project ([18]) identified only a limited number of automatic extraction tools mostly geared to extract technical metadata (e.g.[29],[31]), illustrating the intensive manual labour required in the ingest of digital material into a repository. Subsequently

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 63–74, 2006.

substantial work on descriptive metadata extraction has emerged: e.g. extraction from structured documents have been attempted by MetadataExtractor from University of Waterloo ([27]), Dublin Core Metadata Editor ([11]) and Automatic Metadata Generation (AMG) at the Catholic University of Leuven([2]), and the extraction of bibliographic information from medical articles, based on the detection of contiguous blocks and fuzzy pattern matching, is available from Medical Article Record System (MARS) ([42]) developed at the US National Library of Medicine (NLM)([30]). There have also been previous work on metadata extraction from scientific articles in postscript using a knowledge base of stylistic cues ([19],[20]) and, from the language processing community, there have been results in automatic categorisation of emails ([6],[24]), text categorisation ([39]) and document content summarisation ([43]). Other communities have used image analysis for information extraction from the Internet ([3]), document white space analysis ([9]), graphics recognition in PDF files ([41]), and algorithms for page segmentation ([40]). Despite the wealth of research being conducted, no general tool has yet been developed which can be employed to extract metadata from digital objects of varied types and genres, nor are there dependable extraction tools for the extraction of deeper semantic metadata such as content summary. The research in this paper is motivated by an effort to address this problem by integrating the methods available in the area to create a prototype tool for automatically extracting metadata across many domains at different semantic levels. This would involve:

- constructing a well-structured experimental corpus of one file type (for use in this and future related research);
- summarising and integrating existing research related to automatic metadata extraction;
- determining the limit and scope of metadata that can be extracted and building a prototype descriptive and semantic metadata extraction tool applicable across many domains;
- extending the tool to cover other file types and metadata ; and,
- integrating it with other tools to enable automatic ingest, selection and/or appraisal.

The initial prototype is intended to extract Genre, Author, Title, Date, Identifier, Pagination, Size, Language, Keywords, Composition (e.g. existence and proportion of images, text and links) and Content Summary. In the present paper, we discuss genre classification of digital documents represented in PDF ([32]) as a step towards acquiring the appropriate metadata. The term genre does not always carry a clear meaning. We follow the definition of Kessler ([25]) who refers to genre as “any widely recognised class of texts defined by some common communicative purpose or other functional traits, provided the function is connected to some formal cues or commonalities and that the class is extensible”. For instance, a scientific research article is a theoretical argument or communication of results relating to a scientific subject usually published in a journal and often starting with a title, followed by author, abstract, and body of text,finally ending with a bibliography. One important aspect of genre classification is that it is distinct from subject classification which can coincide over many genres (e.g. a mathematical paper on number theory versus a news article on the proof of Fermat's Last Theorem). The motivation for starting with genre classification is as follows:

- Identifying the genre first will limit the scope of document forms from which to extract other metadata:
  - The search space for further metadata will be reduced; within a single genre, metadata such as author, keywords, identification numbers or references can be expected to appear in a specific style and region.
  - A lot of independent work exists for extraction of metadata within a specific genre which can be combined with a general genre classifier for metadata extraction over many domains (e.g. the papers listed at the beginning of this section).
  - Resources available for extracting further metadata is different for each genre; for instance, research articles unlike newspaper articles come with a list of reference articles closely related to the original article leading to better subject classification.
- Scoping new genres not apparent in the context of conventional libraries is necessary.
- Different institutional collecting policies might focus on digital materials in different genres. Genre classification will support automating the identification, selection, and acquisition of materials in keeping with local collecting guidelines.

We have opted to consider 60 genres (Table 1). This list is not meant to represent a complete spectrum of possible genres; it is meant to be a starting point from which to determine what is possible.

We have focused our attention on different genres represented in PDF files. By limiting the research to one file type we hoped to put a boundary on the problem space. The choice of PDF as the format stems from the fact that

- PDF is a widely used format. Specifically, PDF is a common format for digital objects ingested into digital libraries including eprint services.
- It is a portable format, distributed over many different platforms.
- There are many tools available for conversion to and from other formats.
- It is a versatile format which includes objects of different type (e.g. images, text, links) and different genres (e.g. data structure, fiction, poetry, research article).

In the experiment which follows we worked with a developmental data set collected via the Internet using a random PDF-grabber which

1. selects a random word from a Spell Checker Oriented Word List (from sourceforge.net),
2. searches the Internet using Google for PDF files containing the chosen word,

Table 1. Scope of genres

Groups	Genres
Book	Academic book, Fiction(book), Poetry(book),Other book
Article	Scientific research article, Other research article, Magazine article, News report
Periodicals	Periodicals, Newsletter
Mail	Email, Letter
Thesis	Thesis, Business/Operational report, Technical report, Misc report
List	List,Catalogue
Table	Calendar, Menu, Other table
Proposal	Grant/Project proposal, Legal appeal/proposal/order
Description	Job/Course/Project description, Product/Application description
Minutes	Minutes, Proceedings
Rules	Instruction/Guideline, Regulations
Other	Abstract,Advertisement, Announcement, Appeal/Propaganda, Biography, Chart/Graph,Contract, Drama, Essay, Exam/Worksheet, Fact sheet,Fiction piece, Forms, Forum discussion, Image, Interview, Lecture notes/presentation, Speech transcript, Manual, Memo, Sheet music, Notice, Posters, Programme, Questionnaire, Q & A, Resume/CV, Review, Slides, Poetry piece, Other genre not listed

3. selects a random PDF file from the returned list and places it in a designated folder.

We collected over 4000 documents in this manner. Labelling of this document corpus is still in progress (for genre classification) and is mostly being carried out by one of the authors. Currently 570 are labelled with one of the 60 genres. A significant amount of disagreement is expected in labelling genre even between human labellers; we intend to cross check the labelled data in two ways:

- We will employ others to label the data to determine the level of disagreement between different human labellers; this will enable us to analyse at what level of accuracy the automated system should be expected perform, while also providing us with a gauge to measure the difficulty of labelling individual genres.
- We will gather PDF files which have already been classified into genres as a fresh test data for the classifier; this will also serve as a means of indexing the performance on well-designed classification standards.

Along with the theoretical work of Biber ([7]) on genre structures, there have been a number of studies in automatic genre classification: e.g. Karlgren and Cutting ([23], distinguishing Press, Misc, Non-fiction and Fiction), Kessler et al. ([25], distinguishing Reportage, Fiction, Scitech, Non-fiction, Editorial and Legal; they also attempt to detect the level of readership - which is referred to as Brow - divided into four levels, and make a decision on whether or not the text is a narrative), Santini ([38], distinguishing Conversation, Interview, Public Debate, Planned Speech, Academic prose, Advert, Biography, Instruction, Popular Lore and Reportage), and, Bagdannov and Worring ([4], fine-grained genre classification using first order random graphs modeled on trade journals and brochures found in the Océ Competitive Business Archive) not to mention a recent MSc. dissertation written by Boese ([8], distinguishing ten genres of web documents). There are also related studies in detecting document logical structures ([1]) and clustering documents ([5]). Previous methods can be divided into groups which look at one or more of the following:

- Document image analysis
- Syntactic feature analysis
- Stylistic feature analysis
- Semantic structure analysis
- Domain knowledge analysis

We would eventually like to build a tool which looks at all of these for the 60 genres mentioned (see Table 1). The experiments in this paper however are limited to looking at the first two aspects of seven genres. Only looking at seven genres out of 60 is a significant cut back, but the fact that none of the studies known to us have combined the first two aspects for genre classification and that very few studies looked at the task in the context of PDF files makes the experiments valuable as a report on the first steps to a general process. This paper is not meant to be a conclusive report, but the preliminary findings of an ongoing project and is meant to show the promise of combining very different classifying methods in identifying the genre of a digital document. It is also meant to emphasise the importance of looking at information extraction across genres; genre-specific information extraction methods usually depend heavily on the structures held in common by the documents in the chosen domain; by looking at differences between genres we can determine the variety of structures one might have to resolve in the construction of a general tool.

2 Classifiers

The experiments described in this paper require the implementation of two classifiers:

Image classifier: this classifier depends on features extracted from the PDF document when handled as an image.

- It uses the module pdftoppm from XPDF to extract the first page of the document as an image then employs Python's Image Library (PIL) ([35], [33]) to extract pixel values. This is then sectioned off into ten regions for an examination of the number of non-white pixels. Each region is rated as level 0, 1, 2, 3 (larger number indicating a higher density of non-white space). The result is statistically modelled using the Maximum Entropy principle. The tool used for the modelling is MaxEnt for C++ developed by Zhang Le ([26]).

Language model classifier: this classifier depends on an N-gram model on the level of words, Part-of-Speech tags and Partial Parsing tags.

- N-gram models look at the possibility of word w(N) coming after a string of words W(1), W(2), ..., w(N-1). A popular model is the case when N=3. This model is usually constructed on the word level. In this research we would eventually like to make use of the model on the level of Part-of-Speech (POS) tags (for instance, tags which denote whether a word is a verb, noun or preposition) or Partial Parsing (PP) tags (e.g. noun phrases, verb phrases or prepositional phrases). Initially we only work with the word-level model. This has been modelled by the BOW toolkit developed by Andrew McCallum ([28]). We used the default Naiive Bayes model without a stoplist.

Although the tools for extracting the image and text of the documents used in these classifiers are specific to PDF files, a comparable representation can be extracted in other formats by substituting these tools with corresponding tools for those formats. In the worst-case scenario the process can be approximated by first converting the format to PDF, then using the the same tools; the wide distribution of PDF ensures the existence of a conversion tool for most common formats.

Using the image of a text document in the classification of the document has several advantages:

- it will be possible to extract some basic information about documents without accessing content or violating password protection or copyright;
- more likely to be able to forgo the necessity of substituting language modeling tools when moving between languages, i.e. it maximises the possibility of achieving a language independent tool;
- the classification will not be solely dependent on fussy text processors and language tools (e.g. encoding requirements, problems relating to special characters or line-breaks);
- it can be applied to paper documents digitally imaged (i.e. scanned) for inclusion in digital repositories without heavily relying on accuracy in character recognition.

3 Experiment Design

The experiments in this paper are the first steps towards testing the following hypothesis:

> Hypothesis A: Given a collection of digital documents consisting of several different genres, the set of genres can be partitioned into groups such that the visual characteristics concur and linguistic characteristics differ between documents within a single group, while visual aspects differ between the documents of two distinct groups.

An assumption in the two experiments described here is that PDF documents are one of four categories: Business Report, Minutes, Product/Application Description, Scientific Research Article. This, of course, is a false assumption and limiting the scope in this way changes the meaning of the resulting statistics considerably. However, the contention of this paper is that high level performance on a limited data set combined with a suitable means of accurately narrowing down the candidates to be labelled would achieve the end objective.

Steps for the first experiment

1. take all the PDF documents belonging to the above four genres (70 documents in the current labelled data),
2. randomly select a third of the documents in each genre as training data (27 documents) and the remaining documents as test data (43 documents),
3. train both the image classifier and language model classifier (on the level of words) on the selected training data,
4. examine result.

Steps for the second experiment

1. using the same training and test data as that for the first experiment,
2. allocate the genres to two groups, each group containing two genres: Group I contains business reports and minutes while Group II contains scientific research articles and product descriptions,
3. train the image classifier to differentiate between the two groups and use this to label the test data as documents of Group I or Group II,
4. train two language model classifiers: Classifier I which distinguishes business reports from minutes and Classifier II which labels documents as scientific research articles or product descriptions,
5. take test documents which have been labelled Group I and label them with Classifier I; take test documents which have been labelled Group II and label them with Classifier II,
6. examine result.

The genres to be placed in Group I and Group II were selected by choosing the partition which showed the highest training accuracy for the image classifier.

4 Results

In the evaluation of the results to follow we will use three indices which are considered standard in a classification tasks: accuracy, precision and recall. Let N be the total number of documents in the test data, N_c the number of documents in the test data which are in class C, T the total number of correctly labelled documents in the data independent of the class, T_c the number of true positives for class C (documents correctly labelled as class C), and F_c the number of false positives for class C (documents labelled incorrectly as class C). Accuracy is defined to be $A = \frac{T}{N}$ while precision and recall for each class C is defined to be $P_c = \frac{T_c}{(T_c+F_c)}$ and $R_c = \frac{T_c}{N_c}$ respectively.

The precision and recall for the first and second experiments are given in Table 2 and Table 3.

Table 2. Result for first small experiment

Overall accuracy (Language model only): 77%

Genres	Prec.(%)	Rec.(%)
Business Report	83	50
Sci. Res. Article	88	80
Minutes	64	100
Product Desc.	90	90

Table 3. Result for second small experiment

Overall accuracy(Image and Language model: 87.5 %

Genres	Prec.(%)	Rec(%)
Business Report	83	50
Sci. Res. Article	75	90
Minutes	71	100
Product Desc.	90	100

Although the performance of the language model classifier given in Table 2 is already surprisingly high, this, to a great extent, depends on the four categories chosen. In fact, when the classifier was expanded to include 40 genres, the classifier performed only at an accuracy of approximately 10%. When a different set was employed which included Periodicals, Thesis, Minutes and Instruction/Guideline, the language model performs at an accuracy of 60.34%. It is clear from the two examples that such a high performance can not be expected for any collection of genres.

The image classifier on Group I(Periodicals) and Group II(Thesis, Minutes, Instruction/Guideline) performs at an accuracy of 91.37%. The combination of the two classifiers have not been tested but even in the worst-case scenario, where we assume that the set of mislabelled documents for the two classifiers have no intersection, the combined classifier would still show an increase in overall accuracy of approximately 10%.

The experiments show an increase in the overall accuracy when the language classifier is combined with the image classifier. To gauge the significance of the increase, a statistically valid significance test would be required. The experiments here however are intended not to be conclusive but indicative of the promise underlying the combined system.

5 Conclusion and Further Research

5.1 Intended Extensions

The experiments show that, although there is a lot of confusion visually and linguistically over all 60 genres, subgroups of the genres exhibit statistically well-behaved characteristics. This encourages the search for groups which are similar or different visually or linguistically to further test Hypothesis A. To extend the scenario in the experiment to all the genres the following steps are suggested.

1. randomly select a third of the documents in each genre as training data and the remaining documents as test data,
2. train the image and language model classifier on the resulting and test over all genres,
3. try to re-group genres so that each group contain genres resulting in a high level of cross labelling in the previous experiment,
4. re-train and test.

5.2 Employment of Further Classifiers

Further improvement can be envisioned by integrating more classifiers into the decision process. For instance consider the following classifiers.

Extended image classifier: In the experiments described in this paper the image classifier looked at only the first page of the document. A variation or extension of this classifier to look at different pages of the document or several pages of the document will be necessary for a complete image analysis. This would however involve several decisions: given that documents have different lengths, the optimal number of pages to be used needs to be determined, and we need to examine the best way to combine the information from different pages (e.g. will several pages be considered to be one image; if not, how will the classification of synchronised pages be statistically combined to give a global classification).

Language model classifier on the level of POS and phrases: This is a N-gram language model built on the part-of-speech tags of the undelying text of the document and also on partial chunks resulting from detection of phrases.

Stylo-metric classifier: This classifier takes its cue from positioning of text and image blocks, font styles, font size, length of the document, average sentence lengths and word lengths. This classifier is expected be useful for both genre classification (by distinguishing linguistically similar Thesis and Scientific Research Article by say the length of the document) and other bibliographic data extraction (by detecting which strings are the Title and Author by font style, size and position).

Semantic classifier: This classifier will combine extraction of keywords, subjective or objective noun phrases (e.g. using [36]). This classifier is expected to play an important role in the summarisation stage if not already in the genre classification stage.

Classifier based on external information: When the source information of the document is available, such features as name of the journal, subject or address of the webpage and anchor texts can be gathered for statistical analysis or rule-based classification.

5.3 Labelling More Data

To make any reasonable conclusions with this study, further data needs to be labelled for fresh experiments and also to make up for the lack of training data. Although 60 genres are in play, only 40 genres had more than 3 items in the set and only 27 genres had greater than or equal to 15 items available.

6 Putting It into Context

Assuming we are able build a reasonable extractor for genre, we will move on to implementing the extraction of author, title, date, identifier, keywords, language, summarisations and other compositional properties within each specific genre. After this has been accomplished, we should augment the tool to handle subject classification and to cover other file types.

Once the basic prototype for automatic semantic metadata extraction is tamed into a reasonable shape, we will pass the protype to other colleagues in the Digital Curation Centre ([10]) to be integrated with other tools (e.g. technical metadata extraction tools) and standardised frameworks (e.g. ingest or preservation model) for the development of a larger scale ingest, selection and appraisal application. Eventually, we should be able at least to semi-automate essential processes in this area.

Acknowledgements

This research is being conducted as part of The Digital Curation Centre's (DCC) [10] research programme. The DCC is supported by a grant from the United Kingdom's Joint Information Systems Committee (JISC) [22] and the e-Science Core Programme of the Engineering and Physical Sciences Research Council (EPSRC) [16]. The EPSRC grant (GR/T07374/01) provides the support for the research programme. Additional support for this research comes from the *DELOS: Network of Excellence on Digital Libraries* (G038-507618) funded under the European Commission's IST 6^{th} Framework Programme [12]. The authors would like to thank their DCC colleague Adam Rusbridge whose work on ERPANET's Packaged Object Ingest Project [18] provided a starting point for the current project on automated metadata extraction. We are grateful to the anonymous ECDL reviewers of this paper who provided us with very helpful comments, which enabled us to improve the paper.

Note on website citations: All citations of websites were validated on 29 May 2006.

References

1. Aiello, M., Monz, C., Todoran, L., Worring, M.: Document Understanding for a Broad Class of Documents. International Journal on Document Analysis and Recognition **5(1)** (2002) 1–16.
2. Automatic Metadata Generation: http://www.cs.kuleuven.ac.be/hmdb/amg /documentation.php
3. Arens,A., Blaesius, K. H.: Domain oriented information extraction from the Internet. Proceedings of SPIE Document Recognition and Retrieval 2003 **Vol 5010** (2003) 286.
4. Bagdanov, A. D., Worring, M.: Fine-Grained Document Genre Classification Using First Order Random Graphs. Proceedings of International Conference on Document Analysis and Recognition 2001 (2001) 79.
5. Barbu, E., Heroux, P., Adam, S., Trupin, E.: Clustering Document Images Using a Bag of Symbols Representation. International Conference on Document Analysis and Recognition, (2005) 1216–1220.
6. Bekkerman, R., McCallum, A., Huang, G.: Automatic Categorization of Email into Folders. Benchmark Experiments on Enron and SRI Corpora', CIIR Technical Report, **IR-418** (2004).
7. Biber, D.: Dimensions of Register Variation:a Cross-Linguistic Comparison. Cambridge University Press (1995).
8. Boese, E. S.: Stereotyping the web: genre classification of web documents. Master's thesis, Colorado State University (2005).
9. Breuel, T. M.: An Algorithm for Finding Maximal Whitespace Rectangles at Arbitrary Orientations for Document Layout Analysis. 7th International Conference for Document Analysis and Recognition (ICDAR), 66–70 (2003).
10. Digital Curation Centre: http://www.dcc.ac.uk
11. DC-dot, Dublin Core metadata editor: http://www.ukoln.ac.uk/metadata/dcdot/
12. DELOS Network of Excellence on Digital Libraries: http://www.delos.info/
13. NSF International Projects: http://www.dli2.nsf.gov/ intl.html
14. DELOS/NSF Working Groups: Reference Models for Digital Libraries: Actors and Roles (2003) http://www.dli2.nsf.gov /internationalprojects/ working_group_reports/ actors_final_report.html
15. Dublin Core Initiative: http://dublincore.org/tools/#automaticextraction
16. Engineering and Physical Sciences Research Council: http://www.epsrc.ac.uk/
17. Electronic Resources Preservation Access Network (ERPANET): http:// www.erpanet.org
18. ERPANET: Packaged Object Ingest Project. http://www.erpanet.org/events/ 2003/rome/presentations/ ross_rusbridge_pres.pdf
19. Giuffrida, G., Shek, E. Yang, J.: Knowledge-based Metadata Extraction from PostScript File. Proc. 5th ACM Intl. conf. Digital Libraries (2000) 77–84.
20. Han, H., Giles, L., Manavoglu, E., Zha, H., Zhang, Z., Fox, E. A.: Automatic Document Metadata Extraction using Support Vector Machines. Proc. 3rd ACM/IEEE-CS conf. Digital libraries (2000) 37–48.
21. Hedstrom, M., Ross, S., Ashley, K., Christensen-Dalsgaard, B., Duff, W., Gladney, H., Huc, C., Kenney, A. R., Moore, R., Neuhold, E.: Invest to Save: Report and Recommendations of the NSF-DELOS Working Group on Digital Archiving and Preservation. Report of the European Union DELOS and US National Science Foundation Workgroup on Digital Preservation and Archiving (2003) http://delos-noe.iei.pi.cnr.it/activities/internationalforum/Joint-WGs/digitalarchiving/Digitalarchiving.pdf.

22. Joint Information Systems Committee: http://www.jisc.ac.uk/
23. Karlgren, J. and Cutting, D.: Recognizing Text Genres with Simple Metric using Discriminant Analysis. Proc. 15th conf. Comp. Ling. **Vol 2** (1994) 1071–1075.
24. Ke, S. W., Bowerman, C. Oakes, M. PERC: A Personal Email Classifier. Proceedings of 28th European Conference on Information Retrieval (ECIR 2006) 460–463.
25. Kessler, B., Nunberg, G., Schuetze, H.: Automatic Detection of Text Genre. Proc. 35th Ann. Meeting ACL (1997) 32–38.
26. Zhang Le: Maximum Entropy Toolkit for Python and C++. LGPL license, http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html
27. MetadataExtractor: http://pami-xeon.uwaterloo.ca/TextMiner/ MetadataExtractor.aspx
28. McCallum, A.: Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering. (1998) http://www.cs.cmu.edu/ mccallum/bow/
29. National Archives UK: DROID (Digital Object Identification). http: //www. nationalarchives. gov.uk/ aboutapps/pronom/droid.htm
30. Natinal Library of Medicine US: http://www.nlm.nih.gov/
31. National Library of New Zealand: Metadata Extraction Tool. http://www. natlib. govt.nz/en/whatsnew/4initiatives.html#extraction
32. Adobe Acrobat PDF specification: http://partners.adobe.com/ public/developer/ pdf/index_reference.html
33. Python Imaging Library: http://www.pythonware.com/products/pil/
34. PREMIS (PREservation Metadata: Implementation Strategy) Working Group: http://www.oclc.org/research/projects/pmwg/
35. Python: http://www.python.org
36. Riloff, E., Wiebe, J., and Wilson, T.: Learning Subjective Nouns using Extraction Pattern Bootstrapping. Proc. 7th CoNLL, (2003) 25–32.
37. Ross S and Hedstrom M.: Preservation Research and Sustainable Digital Libraries. International Journal of Digital Libraries (Springer) (2005) DOI: 10.1007/s00799-004-0099-3.
38. Santini, M.: A Shallow Approach To Syntactic Feature Extraction For Genre Classification. Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics (CLUK 04) (2004).
39. Sebastiani F.: 'Machine Learning in Automated Text Categorization', ACM Computing Surveys, **Vol. 34** (2002) 1-47
40. Faisal Shafait, Daniel Keysers, Thomas M. Breuel, "Performance Comparison of Six Algorithms for Page Segmentation", 7th IAPR Workshop on Document Analysis Systems (DAS) (2006).368–379.
41. M. Shao, M. and Futrelle, R.: Graphics Recognition in PDF document. Sixth IAPR International Workshop on Graphics Recognition (GREC2005), 218–227.
42. Thoma,G.: Automating the production of bibliographic records. R&D report of the Communications Engineering Branch, Lister Hill National Center for Biomedical Communications, National Library of Medicine, 2001.
43. Witte, R., Krestel, R. and Bergler, S.: ERSS 2005:Coreference-based Summarization Reloaded. DUC 2005 Document Understanding Workshop, Canada

The Use of Summaries in XML Retrieval

Zoltán Szlávik, Anastasios Tombros, and Mounia Lalmas

Department of Computer Science,
Queen Mary University of London

Abstract. The availability of the logical structure of documents in content-oriented XML retrieval can be beneficial for users of XML retrieval systems. However, research into structured document retrieval has so far not systematically examined how structure can be used to facilitate the search process of users. We investigate how users of an XML retrieval system can be supported in their search process, if at all, through summarisation. To answer this question, an interactive information retrieval system was developed and a study using human searchers was conducted. The results show that searchers actively utilise the provided summaries, and that summary usage varied at different levels of the XML document structure. The results have implications for the design of interactive XML retrieval systems.

1 Introduction

As the *eXtensible Markup Language (XML)* is becoming increasingly used in digital libraries (DL), retrieval engines that allow search within collections of XML documents are being developed. In addition to textual information, XML documents provide a markup that allows the representation of the logical structure of XML documents in content-oriented retrieval. The logical units, called elements, are encoded in a tree-like structure by XML tags. The logical structure allows DL systems to return document portions that may be more relevant to the user than the whole document, e.g. if a searcher wants to read about how Romeo and Juliet met, we do not return the whole play but the actual scene about the meeting. This content-oriented retrieval has received large interest over the last few years, mainly through the INEX initiative [6].

As the number of XML elements is typically large (much larger than that of documents), we believe it is essential to provide users of XML information retrieval systems with overviews of the contents of the retrieved elements. One approach is to use summarisation, which has been shown to be useful in interactive information retrieval (IIR) [9,7,15].

In this paper, we investigate the use of summarisation in XML retrieval in an interactive environment. In interactive XML retrieval, a summary can be associated with each document element returned by the XML retrieval system. Because of the nature of XML documents, users can, in addition to accessing any retrieved element, browse within the document containing that element. One method to allow browsing XML documents is to display the logical structure

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 75–86, 2006.

of the document containing the retrieved elements [13]. This has the benefit of providing (sometimes necessary) context to users when reading an element. Therefore, summaries can also be associated with the other elements of the document, in addition to the returned elements themselves.

The aim of our work is to investigate how users of an XML retrieval system can be supported in their search process, if at all, through summarisation. To answer this question, an interactive information retrieval system was developed and a study using human searchers was conducted.

The paper is organised as follows. In Section 2 we present the background of our work, then we describe the experimental system and methodology that was used in Section 3. The analysis of our data is described in Section 4, which is followed by the conclusions and future work.

2 Background

In recent years, interactive aspects of the IR process have been extensively investigated. Major advances have been made by co-ordinated efforts in the interactive track at TREC. These efforts have been in the context of unstructured documents (e.g. news articles) or in the context of the loosely-defined structure encountered in web pages. XML documents, on the other hand, define a different context, by offering the possibility of navigating within the structure of a single document, or following links to another document part.

The interactive aspect of XML IR has recently been investigated through the interactive track at INEX (iTrack) [13,10,8]. A major result from iTrack 2004 was that searchers did not interact enough with the elements of retrieved XML documents [14]. Searchers seemed to appreciate the logical structure of XML documents as a means of providing context for identifying interesting XML elements within a document, but they did not browse much within XML documents. Tombros et al. suggest that this behaviour may have been due to limitations of the interactive XML IR system used. Among these limitations was that XML element (or document) summarisation capabilities were few, and therefore searchers did not have enough relevance clues to decide which elements to visit [14]. In this paper, we focus on the presentation of the document structure as a hierarchical table of contents, and on the use of summarisation to facilitate the users' search process.

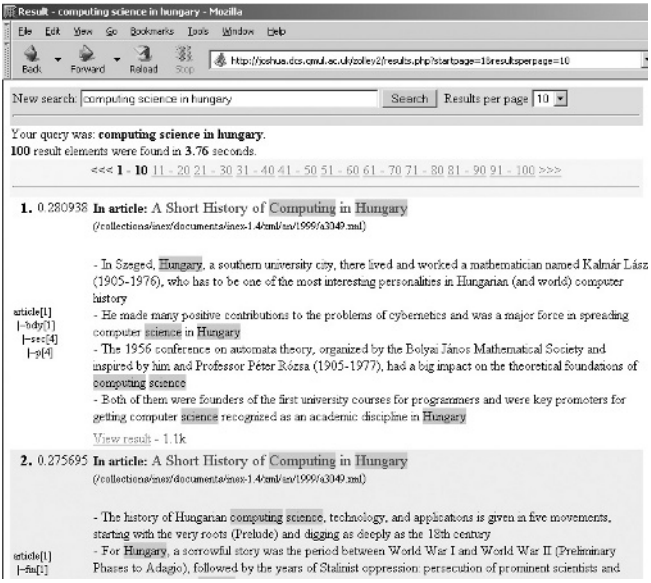
Text summarisation has attracted attention primarily after the information explosion on the Internet; however, significant work was done as early as the 1950's and 1960's. Edmundson proposed extraction methods considering various sentence features, e.g. location, title words [5]. In recent summarisation systems, users' query terms are also considered in generating summaries [15]. Few researchers recently have investigated the summarisation of information available in XML format (e.g. [1,2]). In our work, we considered a simple summarisation algorithm that takes advantage of the sentence location and the query (referred to as query-biased), as our main aim is to study how users "interact" with summaries.

The use of summaries in interactive IR has been shown to be useful for various information seeking tasks in a number of environments such as the web (e.g. [16,4]). However, in the context of interactive XML retrieval, summarisation has not yet been investigated extensively. Our main focus in this paper is to study how searchers behave in an environment that provides them with structural documents, and how they use summaries of document elements that are presented to them. To do so, we created and tested, through user-based studies, an interactive XML retrieval system with XML element summarisation capabilities. We describe the system and the setup of our study in the next section.

3 Experimental Setup

In this section, we describe the system and method that was used in our study. We include only the necessary details for the presentation of the analyis and results reported in this paper. A more detailed description can be found in [12].

User Interface. The user interface is a web based system which passes the query to a retrieval module, processes and displays the retrieved list of elements and shows each of these elements. The system allows users to enter a search query and start the retrieval process by clicking on the search button. The display of the list of retrieved elements is similar to standard search interfaces (Figure 1).


Result - computing science in hungary - Mozilla
File Edit View Go Bookmarks Tools Window Help
Back
Forward
Reload
Stop
http://joshua.dcs.qmul.ac.uk/zoltey2/results.php?startpage=1&resultsperpage=10
Search

New search: computing science in hungary

Results per page 10

Your query was: **computing science in hungary**.
100 result elements were found in **3.76** seconds.

<<< **1 - 10** 11 - 20 21 - 30 31 - 40 41 - 50 51 - 60 61 - 70 71 - 80 81 - 90 91 - 100 >>>

1. 0.280938 **In article:** A Short History of Computing in Hungary
(/collections/inex/documents/inex-1.4/xml/an/1999/a3049.xml)

article[1]
|–bdy[1]
|–sec[4]
|–p[4]

- In Szeged, Hungary, a southern university city, there lived and worked a mathematician named Kalmár Lász (1905-1976), who has to be one of the most interesting personalities in Hungarian (and world) computer history
- He made many positive contributions to the problems of cybernetics and was a major force in spreading computer science in Hungary
- The 1956 conference on automata theory, organized by the Bolyai János Mathematical Society and inspired by him and Professor Péter Rózsa (1905-1977), had a big impact on the theoretical foundations of computing science
- Both of them were founders of the first university courses for programmers and were key promoters for getting computer science recognized as an academic discipline in Hungary

View result - 1.1k

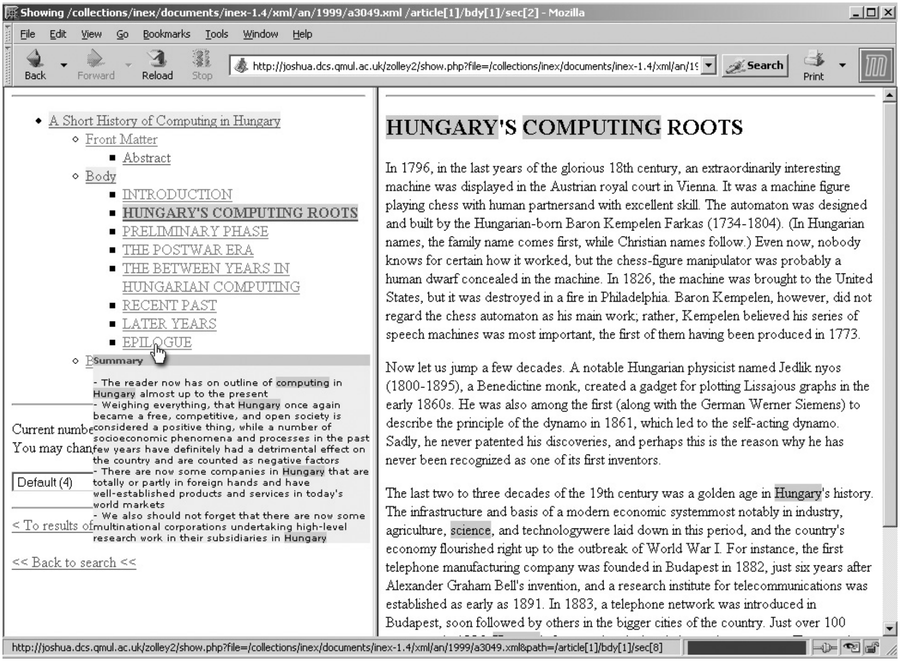
2. 0.275695 **In article:** A Short History of Computing in Hungary
(/collections/inex/documents/inex-1.4/xml/an/1999/a3049.xml)

article[1]
|–fm[1]

- The history of Hungarian computing science, technology, and applications is given in five movements, starting with the very roots (Prelude) and digging as deeply as the 18th century
- For Hungary, a sorrowful story was the period between World War I and World War II (Preliminary Phases to Adagio), followed by the years of Stalinist oppression: persecution of prominent scientists and

Fig. 1. The list of the result elements

Once searchers follow the link to a particular result element, the element is displayed in a new window (Figure 2). The frame on the right shows the content of the target element. The structure is displayed on the left as an automatically generated table of contents (ToC) where each *structural item* is a hyperlink that will show the corresponding XML element in the right window when clicked. For this user study, four levels of structural items were displayed. Level one always refers to the whole article; level two contains the body, front and backmatters; level three usually contains the abstract, sections and appendices; and level four usually means subsections or paragraphs, depending of the inner structure of articles. The number of levels could be changed by searchers. For each item in the ToC, summaries were generated and displayed as 'tool tips', i.e. when users moved the mouse pointer over an item in the ToC, the summary of the target element was shown. Query terms in the summaries, as well as in the displayed documents, were highlighted.



Showing /collections/inex/documents/inex-1.4/xml/an/1999/a3049.xml /article[1]/bdy[1]/sec[2] - Mozilla

http://joshua.dcs.qmul.ac.uk/zolley2/show.php?file=/collections/inex/documents/inex-1.4/xml/an/19

- A Short History of Computing in Hungary
  - Front Matter
    - Abstract
  - Body
    - INTRODUCTION
    - HUNGARY'S COMPUTING ROOTS
    - PRELIMINARY PHASE
    - THE POSTWAR ERA
    - THE BETWEEN YEARS IN HUNGARIAN COMPUTING
    - RECENT PAST
    - LATER YEARS
    - EPILOGUE
  - B

Summary

- The reader now has on outline of computing in Hungary almost up to the present
- Weighing everything, that Hungary once again became a free, competitive, and open society is considered a positive thing, while a number of socioeconomic phenomena and processes in the past few years have definitely had a detrimental effect on the country and are counted as negative factors
- There are now some companies in Hungary that are totally or partly in foreign hands and have well-established products and services in today's world markets
- We also should not forget that there are now some multinational corporations undertaking high-level research work in their subsidiaries in Hungary

Current numbe

You may chan

Default (4)

< To results of

<< Back to search <<

HUNGARY'S COMPUTING ROOTS

In 1796, in the last years of the glorious 18th century, an extraordinarily interesting machine was displayed in the Austrian royal court in Vienna. It was a machine figure playing chess with human partnersand with excellent skill. The automaton was designed and built by the Hungarian-born Baron Kempelen Farkas (1734-1804). (In Hungarian names, the family name comes first, while Christian names follow.) Even now, nobody knows for certain how it worked, but the chess-figure manipulator was probably a human dwarf concealed in the machine. In 1826, the machine was brought to the United States, but it was destroyed in a fire in Philadelphia. Baron Kempelen, however, did not regard the chess automaton as his main work; rather, Kempelen believed his series of speech machines was most important, the first of them having been produced in 1773.

Now let us jump a few decades. A notable Hungarian physicist named Jedlik nyos (1800-1895), a Benedictine monk, created a gadget for plotting Lissajous graphs in the early 1860s. He was also among the first (along with the German Werner Siemens) to describe the principle of the dynamo in 1861, which led to the self-acting dynamo. Sadly, he never patented his discoveries, and perhaps this is the reason why he has never been recognized as one of its first inventors.

The last two to three decades of the 19th century was a golden age in Hungary's history. The infrastructure and basis of a modern economic systemmost notably in industry, agriculture, science, and technologywere laid down in this period, and the country's economy flourished right up to the outbreak of World War I. For instance, the first telephone manufacturing company was founded in Budapest in 1882, just six years after Alexander Graham Bell's invention, and a research institute for telecommunications was established as early as 1891. In 1883, a telephone network was introduced in Budapest, soon followed by others in the bigger cities of the country. Just over 100

http://joshua.dcs.qmul.ac.uk/zolley2/show.php?file=/collections/inex/documents/inex-1.4/xml/an/1999/a3049.xml&path=/article[1]/bdy[1]/sec[8]

Fig. 2. On the left, the structure of the XML document with a summary; on the right, the content of a section element displayed

Summarisation. Summaries were displayed in the result list view for each result element and for the displayed elements in the ToC in element view. Since our aim at this stage of the research was not to develop sophisticated summarisation methods, but to investigate summary usage in XML retrieval, we implemented and used a simple query-biased algorithm. Four sentences with the highest scores were presented as extracts of the source XML elements, in order of appearance in the source element (for further details, see [12]).

Document Collection. The document collection we used was the IEEE collection (version 1.4) which contains 12,107 articles, marked up in XML, of the IEEE Computer Society's publications from 12 magazines and 6 transactions, covering the period of 1995-2002. On average, an article contains 1532 XML nodes and the average depth of a node is 6.9. These properties provided us with a suitably large collection of articles of varying depth of logical structure.

XML Retrieval Engine. The retrieval was based on the HySpirit retrieval framework [11]. To be able to examine the relation between the structural display and the use of summaries, only paragraphs were returned as retrieval results. This strategy ensured that elements deeply nested in a document logical structure were returned, so as to "force" searchers to browse through the structural display on the left panel of Figure 2 (instead of simply scrolling down the right window).

Searchers. Twelve searchers (9 males and 3 females) were recruited for this study. All of them had computer science background as the collection used contained articles from the field of computer science.

Experimental and Control Systems. Two versions of the developed system were used in this study. The control system (S_c) had all the functionalities we described in previous sections, whereas the experimental system (S_e) differed in the display mode of summaries: System S_e displayed summaries only at high levels in the hierarchical structure, i.e. the upper three levels had associated summaries, the fourth level did not. The rationale behind this is that we wanted to see whether searchers' behaviour is affected by the different display. To avoid bias towards the use of the hierarchical structure and summarisation, we employed a blind study, i.e. searchers were not told what the purpose of the study was.

Tasks. Four search tasks were used in the experiments. The tasks described simulated work task situations [3]. We used modified versions of the INEX 2005 ad-hoc track topics which ensured that the tasks were realistic, and that relevant documents could be found in the document collection. Two types of search tasks were chosen. Background type tasks instructed searchers to look for information about a certain topic (e.g. concerns about the CIA and FBI's monitoring of the public) while List type tasks asked searchers to create a list of products that are connected to the topic of their tasks (e.g. a list of speech recognition software). From each group of tasks, searchers could freely choose the one that was more interesting to them. Searchers had a maximum of 20 minutes for each task. This period is defined as a *search session.* Search sessions of the same searcher (i.e. one searcher had two search sessions) are defined and used in this paper as a *user session.*

Search Design. To rule out the fatigue and learning effects that could affect the results, we adopted Latin square design. Participants were randomly assigned into groups of four. Within groups, the system order and the task order were permuted, i.e. each searcher performed two tasks on different systems which involved two different task types. We made an effort to keep situational variables constant, e.g. the same computer settings were used for each subject, the same (and only) experimenter was present, and the place of the experiments was the same.

Data Collected. Two types of events were logged. One type was used to save the users' actions based on their mouse clicks (e.g. when users clicked on the 'search' button, or opened an element for reading). The other type corresponds to the summary-viewing actions of users, i.e. we logged whenever a summary was displayed (users moved the mouse pointer over an item in the ToC).

During the analysis of summary log files, calculated summary-viewing times that were shorter than half a second or longer than twenty seconds were discarded, because the former probably corresponds to a quick mouse move (without users having read the summary), and the latter may have recorded user actions when the keyboard only was used (e.g. opening another window by pressing CTRL+N).

4 Analysis

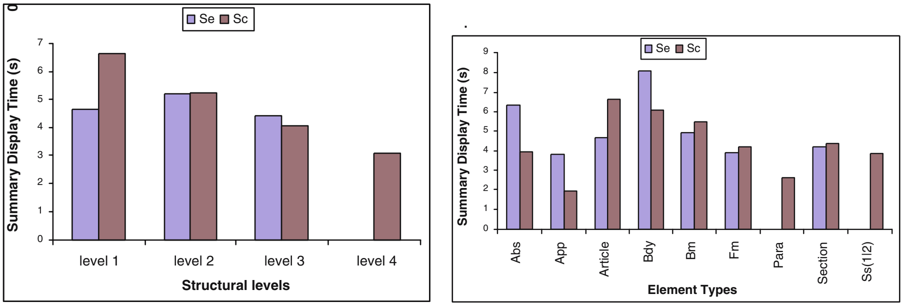
In this section, the analysis of the recorded log files is described. To investigate whether summarisation can be effectively used in interactive XML retrieval, we formed four groups of research questions. The first group (Section 4.1) is about summary reading times. The second group (Section 4.2) is about the number of summaries searchers read in their search sessions. Section 4.3 investigates the relation between summary reading times and number of summaries read (the third group). The fourth group (Section 4.4) looks into the relation between the multi-level XML retrieval and traditional retrieval.

4.1 Summary Times

In this section, we examine how long searchers read an average summary; whether there are differences in reading times for summaries that are associated with elements at various structural levels and element types; and whether the average summary reading time changes when summaries are not shown at all structural levels in the ToC.

Taking into account both systems S_e and S_c, an average summary was displayed for 4.24 seconds with a standard deviation of 3.9. The longest viewed summary was displayed for 19.57s, while the shortest accounted summary was viewed for 0.51s.

Figure 3a shows the distribution of summary display times by structural levels for each system. Display times of S_c tend to be shorter when users read summaries of deeper, i.e. shorter elements, although the length of summaries were the same (i.e. four sentences). For S_e, times are more balanced. This indicates that if there are summaries for more levels and the lowest level is very short (sometimes these paragraphs are as short as the summary itself), people trust summaries of larger, i.e. higher, elements more. If the difference in size between the deepest and highest elements is not so big, times are more balanced.



(a) (b)

Fig. 3. Summary times by structural levels (a) and XML element types (b)

Figure 3b shows the display time distribution by XML element types (tags). We can see, for example, that the *bdy* (body) element has high summary viewing times; this is the element that contains all the main text of the article. We can also see that paragraphs (*para*) and subsections (*ss*1 and *ss*2) have low summary reading times for S_c and, obviously, none for S_e (as they are not displayed at these levels). These three element types appear on the lowest, i.e. fourth, structural level.

We compared the two systems (S_e and S_c) to find out whether significant differences in summary reading times can be found. The comparison of the overall summary-viewing times showed significant difference between S_e and S_c, i.e. the average summary viewing time for system S_e (4.58s) is significantly higher than that of system S_c (3.98s). To examine where this difference comes from, we compared the two systems by tag types (e.g. whether summary reading times for sections are different for the two systems). However, we did not find significant differences at comparable tag types[1]. We also compared the two systems with respect to structural levels (e.g. whether average summary reading time at level one is significantly different for the two systems). We did not find significant difference for level one (article), two (body, front and back matters) and three (abstract, sections, appendix) elements.

To sum up, our results showed that users of system S_e read summaries 0.5s longer than that of system S_c. However, we could not find significant difference at levels or element types between the two systems. An interpretation of this result is that since S_e searchers had less available summaries to examine, they were less confused and overloaded by the information available and could take their time reading a particular summary.

4.2 Number of Summaries Read

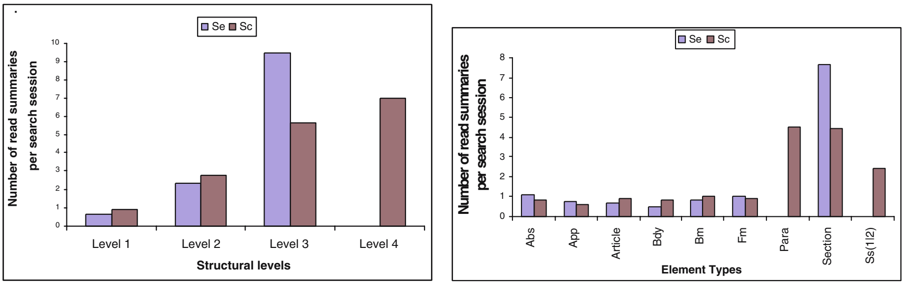
This section looks into the number of summaries that were read by searchers. We first examine the average number of summaries seen by users in a search session,

[1] Tag types for which summaries were not displayed for any of the systems were not compared as one of the sample groups would contain zero samples.

and then we look into the distributions of the number of read summaries at different structural levels and element types. Differences between the two systems with respect to the number of read summaries are also discussed in this section.

Considering both systems together, an average user read 14.42 summaries in a search session (20 minutes long), with a standard deviation of 10.77. This shows a considerable difference in user behaviour regarding summary reading. The least active summary reader read only one summary in a search session, while the most active saw 52 summaries for at least half a second.

Figure 4a shows that the deeper we are in the structure of the ToC, the more summaries are read, on average, in a search session. This is consistent with the nature of XML, and all tree-like structures: the deeper we are in a tree, the more elements are available on that level. However, our data shows that the difference between the two systems is not only based on this structural property, because when only three levels of summaries were displayed, reading of third level summaries (usually summaries of sections) showed higher activity than when four levels of summaries were displayed, i.e. the third level seems to be more interesting than the first and second.



(a) **(b)**

Fig. 4. Number of read summaries per search sessions, by structural levels (a) and XML element types (b)

The next step is to find out whether this interest is only at these deeper levels, or connected to some element types. Contents of the same element types are supposed to have the same amount and kind of information, e.g. paragraphs are a few sentences long; front matters usually contain author, title information and the abstract of the paper. Our log analysis shows that summaries of sections, subsections and paragraphs are those most read (Figure 4b), although users take less time to read them (see previous section). Other tag types are less promising to users according to their summary usage. We can also see in Figure 4b that when paragraph and subsection summaries are not available (S_e), section summary reading increases dramatically. We interpret these results as indication that for the IEEE collection, sections, that appear mostly at level three, are the most promising elements to look at when answering an average query.

The comparison of the overall number of viewed summaries showed that an average user of system S_e read 12.5 summaries per search session, and of system S_c 16.33 summaries per session. In other words, our test persons read more summaries where more summaries were available. However, this difference is not statistically significant.

We compared the two systems using the same categories (i.e. tag types and levels) as previously for summary reading times. T-tests did not show significant differences at comparable levels and element types between S_e and S_c in number of read summaries.

4.3 Reading Times vs. Number of Read Summaries

In this section, we examine the relationships between the data and findings of the previous two sections. One question we are looking into is whether searchers with higher summary reading times read less summaries in a search session.

Users of system S_e read less summaries than those who used system S_c. This is in accordance that they had less summaries available. However, users of system S_e also read summaries for longer. This shows that if there are less available summaries, users can focus more on one particular summary, and vice versa, if there are many summaries to view, reading can become superficial.

Considering both systems and tag types, we found negative correlation between the summary reading time and the number of read summaries. In other words, it is true for users of both systems that the more summaries they read on a particular level the shorter the corresponding reading times are. However, this is only an indication as the correlation coefficient (-0.5) does not indicate significance. Also, since the number of summaries read increases when going deeper in the structure, we view this as an indication that, for searchers, summaries of higher level elements are more indicative to the contents of the corresponding elements than those of lower, and also shorter, elements.

4.4 Usage of the ToC and Article Level

XML retrieval has the advantage of breaking down a document into smaller elements according to the document's logical structure. We investigated whether searchers take advantage of this structure: do they click on items in the ToC, do they use the article (unstructured) level of a document, and how frequently, do they alternate between full article and smaller XML element views?

Regarding the usage of the XML structure in terms of the ToC, 58.16% of the displayed elements were results of at least "second" clicks, i.e. more than half of the elements were displayed by clicking on an element in the ToC. This shows that searchers actively used the ToC provided (unlike those in [14]), and that they used the logical structure of the documents by browsing within the ToC.

The log files show that only 25% of the searchers clicked on article elements to access the whole document, and none of these searchers clicked on an article type link more than three times in a search session. The distribution of viewing whole articles did not depend on the system, i.e. we observed three article clicks for each system. This result follows naturally, since the display of the article level was not different in S_e and S_c.

Article level clicks show that articles were only 3.56% of all the displayed elements. This may be misleading as the retrieval system did not return article elements in the result list. We therefore compared article usage to elements that were displayed when users were already in the document view, i.e. we excluded elements that were shown right after a searcher clicked on a link in the result list. The updated number shows that article elements were displayed in 6.12% of these clicks. This suggests that searchers of an XML retrieval system do use the structure available in terms of the ToC, and although it was the first time they had used an XML retrieval system, they did not simply prefer to see the whole document as they were accustomed to.

Our results from the previous sections suggest that searchers still want to have access to, and use, the full-article level. For example, searchers read summaries of articles and read them for longer but, they did not necessarily want to use the full-articles directly, i.e. looking at the full-article summary may be enough to decide whether reading any part of the article is worthwhile.

5 Conclusions and Future Work

In this paper, we presented a study of an interactive XML retrieval system that uses summarisation to help users in navigation within XML documents. The system takes advantage of the logical structure of an XML document by presenting the structure in the form of a table of contents (ToC). In this ToC, summaries of corresponding XML elements were generated and displayed.

Searchers in our study did indeed use the provided structure actively and did not only use the whole article in order to identify relevant content. In addition, searchers made good use of the XML element summaries, by spending a significant amount of time reading these summaries. This implies that our system, by the use of summarisation, facilitated browsing in the ToC level more than that at INEX 2004 interactive track [13].

Regarding the use of element summaries, in our study searchers tended to read more summaries that were associated with elements at lower levels in the structure (e.g. summaries of paragraphs), and at the same time summaries of lower elements were read for a shorter period of time. Our results also suggest that if more summaries are made available, searchers tend to read more summaries in a search session, but for shorter time.

In our experiment, the ToC display and summary presentation were highly connected (i.e. no summary can be displayed without a corresponding item in the ToC). Based on the close relation between them, for such an XML retrieval system it is important to find the appropriate ToC, and summary, presentation. If the ToC is too deep, searchers may lose focus as the reading of many summaries and short reading times at low levels indicated. Nevertheless, if the ToC is not detailed enough, users may lose possibly good links to relevant elements. Our results suggest, that for the used collection, a one or two-level ToC (containing reference to the whole article, body, front and back matter) would be probably too shallow, while displaying the full fourth level (normally to paragraph-level) is sometimes too deep.

We view our results as having implications for the design of interactive XML IR systems that support searchers by providing element summaries and structural information. One implication of the results is that the display of the ToC in XML IR systems needs to be carefully chosen (see previous paragraph). Our results also showed that summarisation can be effectively used in XML retrieval. A further implication of the results is that XML retrieval systems should consider the logical structure of the document for summary generation, as searchers do not use summaries in the same way at all levels of the structure.

Based on the outcomes of this study, our future work includes the development of an improved interactive XML retrieval system that includes adaptive generation of summaries at the ToC level, where the structural position and estimated relevance of the element to be summarised will also be considered (some initial work is done in [2]). We also plan to take into account structural IR search task types (e.g. fetch and browse), that are currently being investigated at INEX [6]. The aim of the fetch and browse retrieval strategy is to first identify relevant documents (the fetching phase), and then to identify the most relevant elements within the fetched articles (the browsing phase). We believe that summarisation can be particularly helpful in the browsing phase, where finding relevant elements within a document is required.

Acknowledgments

This work was partly funded by the DELOS Network of Excellence in Digital Libraries, to which we are very grateful.

References

1. A. Alam, A. Kumar, M. Nakamura, A. F. Rahman, Y. Tarnikova, and C. Wilcox. Structured and unstructured document summarization: Design of a commercial summarizer using lexical chains. In *ICDAR*, pages 1147–1152. IEEE Computer Society, 2003.
2. M. Amini, A. Tombros, N. Usunier, M. Lalmas, and P. Gallinari. Learning to summarise XML documents by combining content and structure features (poster). In *CIKM'05*, Bremen, Germany, October 2005.
3. P. Borlund. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3), 2003.
4. S. Dumais, E. Cutrell, and H. Chen. Optimizing search by showing results in context. In *CHI '01: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 277–284, New York, NY, USA, 2001. ACM Press.
5. H. P. Edmundson. New methods in automatic extracting. *J. ACM*, 16(2):264–285, 1969.
6. N. Fuhr, M. Lalmas, S. Malik, and Z. Szlávik, editors. *Advances in XML Information Retrieval: Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004*, volume 3493. Springer-Verlag GmbH, may 2005.

7. J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell. Summarizing text documents: sentence selection and evaluation metrics. In *SIGIR'99*, pages 121–128. ACM Press, 1999.
8. H. Kim and H. Son. Interactive searching behavior with structured xml documents. In Fuhr et al. [6], pages 424–436.
9. J. Kupiec, J. Pedersen, and F. Chen. A trainable document summarizer. In *SIGIR'95*, pages 68–73. ACM Press, 1995.
10. N. Pharo and R. Nordlie. Context matters: An analysis of assessments of XML documents. In Fabio Crestani and Ian Ruthven, editors, *CoLIS*, volume 3507 of *Lecture Notes in Computer Science*, pages 238–248. Springer, 2005.
11. T. Rölleke, R. Lübeck, and G. Kazai. The hyspirit retrieval platform. In *SIGIR'01*, page 454, New York, NY, USA, 2001. ACM Press.
12. Z. Szlávik, A. Tombros, and M. Lalmas. Investigating the use of summarisation for interactive XML retrieval. In *Proceedings of the 21st ACM Symposium on Applied Computing, Information Access and Retrieval Track (SAC-IARS'06), to appear*, pages 1068–1072, 2006.
13. A. Tombros, B. Larsen, and S. Malik. The interactive track at INEX 2004. In Fuhr et al. [6].
14. A. Tombros, S. Malik, and B. Larsen. Report on the INEX 2004 interactive track. *ACM SIGIR Forum*, 39(1):43–49, June 2005.
15. A. Tombros and M. Sanderson. Advantages of query biased summaries in information retrieval. In *SIGIR'98*, pages 2–10. ACM Press, 1998.
16. R. W. White, J. M. Jose, and I. Ruthven. A task-oriented study on the influencing effects of query-biased summarisation in web searching. *Inf. Process. Manage.*, 39(5):707–733, 2003.

An Enhanced Search Interface for Information Discovery from Digital Libraries

Georgia Koutrika[1],* and Alkis Simitsis[2],**

[1] University of Athens,
Department of Computer Science,
Athens, Greece
koutrika@di.uoa.gr
[2] National Technical University of Athens,
Department of Electrical and Computer Engineering,
Athens, Greece
asimi@dbnet.ece.ntua.gr

Abstract. Libraries, museums, and other organizations make their electronic contents available to a growing number of users on the Web. A large fraction of the information published is stored in structured or semi-structured form. However, most users have no specific knowledge of schemas or structured query languages for accessing information stored in (relational or XML) databases. Under these circumstances, the need for facilitating access to information stored in databases becomes increasingly more important. Précis queries are free-form queries that instead of simply locating and connecting values in tables, they also consider information around these values that may be related to them. Therefore, the answer to a précis query might also contain information found in other parts of the database. In this paper, we describe a précis query answering prototype system that generates personalized presentation of short factual information précis in response to keyword queries.

1 Introduction

Emergence of the World Wide Web has given the opportunity to libraries, museums, and other organizations to make their electronic contents available to a growing number of users on the Web. A large fraction of that information is stored in structured or semi-structured form. However, most users have no specific knowledge of schemas or (semi-)structured query languages for accessing information stored in (relational or XML) databases. Under these circumstances, the need for facilitating access to information stored in databases becomes increasingly more important.

Towards that direction, existing efforts have mainly focused on facilitating querying over structured data proposing either handling natural language queries [2, 14, 17] or free-form queries [1, 18]. However, end users want to achieve their goals

* This work is partially supported by the Information Society Technologies (IST) Program of the European Commission as part of the DELOS Network of Excellence on Digital Libraries (Contract G038-507618).

** This work is co-funded by the European Social Fund (75%) and National Resources (25%) - Operational Program for Educational and Vocational Training II (EPEAEK II) and particularly the Program PYTHAGORAS.

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 87 – 98, 2006.

with a minimum of cognitive load and a maximum of enjoyment [12]. In addition, they often have very vague information needs or know a few buzzwords. Therefore, the usefulness of keyword-based queries, especially compared to a natural language approach in the presence of complex queries, has been acknowledged [26].

Consider a digital collection of art works made available to people on the Web. A user browses the contents of this collection with the purpose of learning about "Michelangelo". If this need is expressed as a free-form query, then existing keyword searching approaches focus on finding and possibly interconnecting entities that contain the query terms, thus they would return an answer as brief as "*Michelangelo: painter, sculptor*". This answer conveys little information to the user and more importantly does not help or encourage him in searching or learning more about "Michelangelo". On the other hand, a more complete answer containing, for instance, biographical data and information about this painter's work would be more meaningful and useful instead. This could be in the form of the following précis:

> "*Michelangelo (March 6, 1475 - February 18, 1564) was born in Caprese, Tuscany, Italy. As a painter, Michelangelo's work includes Holy Family of the Tribune (1506), The Last Judgment (1541), The Martyrdom of St. Peter (1550). As a sculptor Michelangelo's work includes Pieta (1500), David (1504).*"

A précis is often what one expects in order to satisfy an information need expressed as a question or as a starting point towards that direction. Based on the above, support of free-form queries over databases and generation of answers in the form of a précis comprises an advanced searching paradigm helping users to gain insight into the contents of a database. A précis may be incomplete in many ways; for example, the abovementioned précis of "Michelangelo" includes a non-exhaustive list of his works. Nevertheless, it provides sufficient information to help someone learn about Michelangelo and identify new keywords for further searching. For example, the user may decide to explicitly issue a new query about "*David*" or implicitly by following underlined topics (hyperlinks) to pages containing relevant information.

In the spirit of the above, recently, précis queries have been proposed [11]. These are free-form queries that instead of simply locating and connecting values in tables, they also consider information around these values that may be related to them. Therefore, the answer to a précis query might also contain information found in other parts of the database, e.g., frescos created by Michelangelo. This information needs to be "assembled" -in perhaps unforeseen ways- by joining tuples from multiple relations. Consequently, the answer to a précis query is a whole new database, a logical database subset, derived from the original database compared to flattened out results returned by other approaches. This subset is useful in many cases and provides to the user much greater insight into the original data.

The work that we describe in this paper focuses on design and implementation issues of a précis query answering prototype with the following characteristics:

- Support of a keyword-based search interface for accessing the contents of the underlying collection.
- Generation of a logical subset of the database that answers the query, which contains not only items directly related to the query selections but also items implicitly related to them in various ways.

- Personalization of the logical subset generated and hence the précis returned according to the needs and preferences of the user as a member of a group of users.
- Translation of the structured output of a précis query into a synthesis of results. The output is an English presentation of short factual information précis.

Outline. Section 2 discusses related work. Section 3 describes the general framework of précis queries. Section 4 presents the design and implementation of our prototype system, and Section 5 concludes our results with a prospect to the future.

2 Related Work

The need for free-form queries has been early recognized in the context of databases [18]. With the advent of the World Wide Web, the idea has been revisited. Several research efforts have emerged for keyword searching over relational [1, 3, 8, 13] and XML data [5, 6, 9]. Oracle 9i Text [19], Microsoft SQL Server [16] and IBM DB2 Text Information Extender [10] create full text indexes on text attributes of relations and then perform keyword queries.

Existing keyword searching approaches focus on finding and possibly interconnecting tuples in relations that contain the query terms. For example, the answer for “Michelangelo” would be in the form of relation-attribute pair, such as (Artist, Name). In many practical scenarios, this answer conveys little information about “Michelangelo”. A more complete answer containing, for instance, information about this artist's works would be more useful. In the spirit of the above, recently, précis queries have been proposed [11]. The answer to a précis query is a whole new database, a logical database subset, derived from the original database. Logical database subsets are useful in many cases. However, naïve users would rather prefer a friendly representation of the information contained in a logical subset, without necessarily understanding its relational character. In earlier work [11], the importance of such representation constructed based on information conveyed by the database graph, has been suggested. A method for generating an English presentation of the information contained in a logical subset as a synthesis of simple SPO sentences has been proposed [21]. The process resembles those involved in handling natural language queries over relational databases in that they both involve some amount of additional predefinitions for the meanings represented by relations, attributes and primary-to-foreign key joins. However, natural language query processing is more complex, since it has to handle ambiguities in natural language syntax and semantics whereas this approach uses well defined templates to rephrase relations and tuples.

The problem of facilitating the naïve user has been thoroughly discussed in the field of natural language processing (NLP). For the last couple of decades, several works are presented concerning NL Querying [26, 15], NL and Schema Design [23, 14, 4], NL and DB interfaces [17, 2], and Question Answering [25, 22]. Related literature on NL and databases, has focused on totally different issues such as the interpretation of users’ phrasal questions to a database language, e.g., SQL, or to the automatic database design, e.g., with the usage of ontologies [24]. There exist some recent efforts that use phrasal patterns or question templates to facilitate the answering procedure [17, 22]. Moreover, these works produce pre-specified answers, where only the values in the patterns change. This is in contrast to précis queries, which construct logical subsets on demand and use templates and constructs of sentences defined on the constructs of the database graph, thus generating dynamic answers. This characteristic of précis queries also enables template multi-utilization.

In this paper, we built upon the ideas of [11, 20, 21] and we describe the design and implementation of a system that supports précis queries for different user groups.

3 The Précis Query Framework

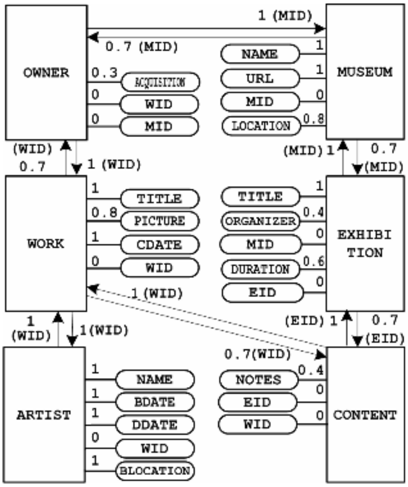
The purpose of this section is to provide background information on précis queries.

Preliminaries. We consider the *database schema graph* $\mathbf{G}(\mathbf{V}, \mathbf{E})$ as a directed graph corresponding to a database schema $\mathtt{D}$. There are two types of nodes in $\mathbf{V}$: (a) *relation nodes,* $\mathbf{R}$, one for each relation in the schema; and (b) *attribute nodes*, $\mathbf{A}$, one for each attribute of each relation in the schema. Likewise, edges in $\mathbf{E}$ are the following: (a) *projection edges*, $\mathbf{\Pi}$, each one connects an attribute node with its container relation node, representing the possible projection of the attribute in the system's answer; and (b) *join edges*, $\mathbf{J}$, from a relation node to another relation node, representing a potential join between these relations. These could be joins that arise naturally due to foreign key constraints, but could also be other joins that are meaningful to a domain expert. Joins are directed for reasons explained later. Therefore, a database graph is a directed graph $\mathbf{G}(\mathbf{V}, \mathbf{E})$, where: $\mathbf{V} = \mathbf{R} \cup \mathbf{A}$, and $\mathbf{E} = \mathbf{\Pi} \cup \mathbf{J}$.

A *weight*, $\mathtt{w}$, is assigned to each edge of the graph $\mathbf{G}$. This is a real number in the range $[0, 1]$, and represents the significance of the bond between the corresponding nodes. Weight equal to 1 expresses strong relationship; in other words, if one node of the edge appears in an answer, then the edge should be taken into account making the other node appear as well. If a weight equals to 0, occurrence of one node of the edge in an answer does not imply occurrence of the other node. Based on the above, two relation nodes could be connected through two different join edges, in the two possible directions, between the same pair of attributes, but carrying different weights. For simplicity, we assume that there is at most one directed edge from one node to the same destination node.

A directed path between two relation nodes, comprising adjacent join edges, represents the "implicit" join between these relations. Similarly, a directed path between a relation node and an attribute node, comprising a set of adjacent join edges and a projection edge represents the "implicit" projection of the attribute on this relation. The weight of a path is a function of the weights of constituent edges, which should satisfy the condition that the estimated weight should decrease as the length of the path increases, based on human intuition and cognitive evidence. In our system, we have considered the product of weights over a path.

Logical Database Subsets. Consider a database $\mathtt{D}$ properly annotated with a set of weights and a *précis query* $\mathtt{Q}$, which is a set of tokens, i.e. $\mathtt{Q}=\{k_1, k_2, \ldots, k_m\}$. We define as *initial relation* any database relation that contains at least one tuple in which one or more query tokens have been found. A tuple containing at least one query token is called *initial tuple*.



ARTIST(wid,name,birth_date,
death_date,birth_location)

CONTENT (wid,eid,notes)

EXHIBITION(eid,mid,title,
duration,organizer)

MUSEUM(mid,name,url,location)

OWNER (wid,mid,acquisition)

WORK(wid,title,creation_date,
picture)

Fig. 1. An example database graph

A *logical database subset* D′ of D satisfies the following:

- The set of relation names in D′ is a subset of that in the original database D.
- For each relation R_i' in the result D′, its set of attributes in D′ is a subset of its set of attributes in D.
- For each relation R_i' in the result D′, the set of its tuples is a subset of the set of tuples in the original relation R_i in D (when projected on the set of attributes that are present in the result).

The result of applying query Q on a database D given a set of constraints C is a logical database subset D′ of D, such that D′ contains initial tuples for Q and any other tuple in D that can be transitively reached by joins on D starting from *some* initial tuple, subject to the constraints C [11]. Possible constraints could be the maximum number of attributes in D′, the minimum weight of paths in D′, the maximum number of tuples in D′ and so forth. Using different constraints and weights on the edges of the database graph allows generating different answers for the same query.

Précis Patterns. Weights and constraints may be provided in different ways. They may be set by the user at query time using an appropriate user interface. This option is attractive in many cases since it enables interactive exploration of the contents of a database. This bears a resemblance to query refinement in keyword searches. In case of précis queries, the user may explore different regions of the database starting, for example, from those containing objects closely related to the topic of a query and progressively expanding to parts of the database containing objects more loosely related to it. Although this approach is quite elegant, the user should spend some time on a procedure that may not always seem relevant to his need for a certain answer. Thus, weights and criteria may be pre-specified by a designer, or stored as part of a profile corresponding to a user or a group of users. In particular, in our framework, we have adopted the use of patterns of logical subsets corresponding to different queries or groups of users, which are stored in the system [20]. For instance, different patterns would be used to capture preferences of art reviewers and art fans.

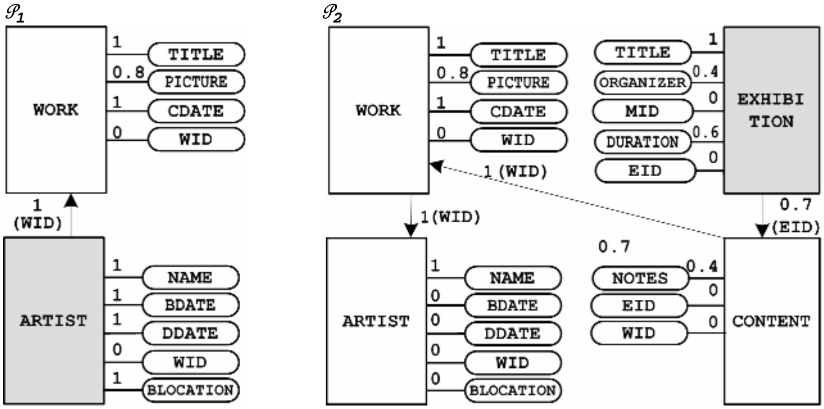


Fig. 2. Example précis patterns

Formally, given the database schema graph G of a database D, a *précis pattern* is a directed rooted tree $\mathcal{P}$(V,E) on top of G annotated with a set of weights. Given a query Q over database D, a précis pattern $\mathcal{P}$(V,E) is *applicable* to Q, if its root relation coincides with an initial relation for Q. The result of applying query Q on a database D given an applicable pattern $\mathcal{P}$ is a logical database subset D′ of D, such that:

– The set of relation names in D′ is a subset of that in $\mathcal{P}$.

– For each relation R_i′ in the result D′, its set of attributes in D′ is a subset of its set of attributes in $\mathcal{P}$.

– For each relation R_i′ in the result D′, the set of its tuples is a subset of the set of tuples in the original relation R_i in D (when projected on the set of attributes that are present in the result).

In order to produce the logical database subset D′, a précis pattern $\mathcal{P}$ is enriched with tuples extracted from the database based on constraints, such as the maximum number of attributes in D′, the maximum number of tuples in D′ and so on.

Example. Consider the database graph presented in Fig. 1. Observe the two directed edges between WORK and OWNER. Works and owners are related but one may consider that owners are more dependent on works than the other way around. In other words, an answer regarding an owner should always contain information about related works, while an answer regarding a work may not necessarily contain information about its owner. For this reason, the weight of the edge from OWNER to WORK is set to 1, while the weight of the edge from WORK to OWNER is 0.7. Précis patterns corresponding to different queries and/or groups of users may be stored in the system. In Fig. 2, patterns $\mathcal{P}_1$ and $\mathcal{P}_2$ correspond to different queries, regarding artists and exhibitions, respectively (the initial relation in each pattern is shown in grey).

4 System Architecture

In this section, we describe the architecture of a prototype précis query answering system, depicted in Fig. 3.

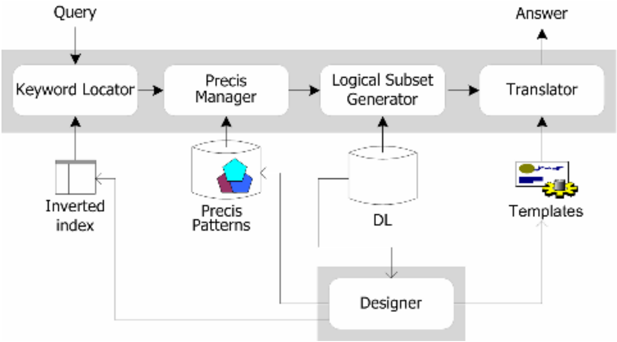


Fig. 3. System Architecture

Each time a user poses a question, the system finds the initial relations that match this query, i.e. database relations containing at least one tuple in which one or more query tokens have been found (Keyword Locator). Then, it determines the database part that contains information related to the query; for this purpose, it searches in a repository of précis patterns to extract an appropriate one (Précis Manager). If an appropriate pattern is not found, then a new one is created and registered in the repository. Next, this précis pattern is enriched with tuples extracted from the database according to the query keywords, in order to produce the logical database subset (Logical Subset Generator). Finally, an answer in the form of a précis is returned to the user (Translator). The creation and maintenance of the inverted index, patterns and templates is controlled through a Designer component. In what follows, we discuss in detail the design and implementation of these components.

Designer Interface. This module provides the necessary functionality that allows a designer to create and maintain the knowledge required for the system to operate, i.e.:

– *inverted index*: with a click of a button, the designer may create or drop the inverted index for a relational database.
– *templates*: through a graphical representation of a database schema graph, the designer may define templates to be used by the Translator.
– *user groups*: the designer may create pre-specified groups of users. Then, when a new user registers in the system, he may choose the group he belongs to.
– *patterns*: through a graphical representation of a database schema graph, the designer may define précis patterns targeting different groups of users and different types of queries for a specific domain. These are stored in a repository.

Manual creation of patterns and user groups assumes good domain and application knowledge and understanding. For instance, the pattern corresponding to a query about art works would probably contain the title and creation date of art works along with the names of the artists that created them and museums that own them; whereas the pattern corresponding to a query about artists would most likely contain detailed information about artists such as name, date and location of birth, and date of death along with titles of works an artist has created. Furthermore, different users or groups of users, e.g., art reviewers vs. art fans, would be interested in different logical subsets for the same query. We envision that the system could learn and adapt précis patterns for different users or groups of users by using logs of past queries or by means of social tagging by large numbers of users. Then, the only work a designer would have to do would be the creation of templates.

Keyword Locator. When a user submits a précis query $Q=\{k_1, k_2, \ldots, k_m\}$, the system finds the initial relations that match this query, i.e. database relations containing at least one tuple in which one or more query tokens have been found. For this purpose, an inverted index has been built, which associates each keyword that appears in the database with a list of occurrences of the keyword. Modern RDBMS' provide facilities for constructing full text indices on single attributes of relations (e.g., Oracle9i Text). In our approach, we chose to create our own inverted index, basically due to the following reasons: (a) a keyword may be found in more than one tuple and attribute of a single relation and in more than one relation; and (b) we consider keywords of other data types as well, such as date and number.

At its current version, the system considers that query keywords are connected with the logical operator or. Keywords enclosed in quotation marks, e.g., "Leonardo da Vinci", are considered as one keyword that must be found in the same tuple. This means that the user can issue queries such as "Michelangelo" or "Leonardo da Vinci", but not queries such as "Michelangelo" and "Leonardo da Vinci", which would essentially ask about the connection between these two entities/people. We are currently working on supporting more complex queries involving operators and and not.

Based on the above, given a user query, Keyword Locator consults the inverted index, and returns for each term k_i in Q, a list of all *initial relations*, i.e. $k_i \rightarrow \{R_j\}$, $\forall k_i$ in Q. (If no tuples contain the query tokens, then an empty answer is returned.)

Précis Manager. Précis Manager determines the schema of the logical database subset, i.e. the database part that contains information related to the query. This should involve initial relations and relations around them containing relevant information. The schema of the subset that should be extracted from a database given a précis query may vary depending on the type of the query issued and the user issuing the query. Patterns of logical subsets corresponding to different queries or groups of users are stored in the system. For instance, different patterns would be used to capture preferences of art reviewers and fans.

Each time an individual poses a question, Précis Manager searches into the repository of précis patterns to extract those that are appropriate for the situation. If users are categorized into groups, then this module examines only patterns assigned to the group the active user belongs to. Based on the initial relations identified for query Q, one or more applicable patterns may be identified. Recall that a précis pattern $\mathcal{P}(\mathbf{V},\mathbf{E})$ is applicable to Q, if its root relation coincides with an initial relation for Q. For instance, given a query on "David", a pattern may correspond to artists ("Michelangelo") and another to owners ("Accademia di Belle Arti, Florence, Italy").

If none is returned for a certain initial relation, then the request is propagated to a Schema Generator. This module is responsible for finding which part of the database schema may contain information related to Q. The output of this step is the schema D' of a logical database subset comprised of: (a) relations that contain the tokens of Q; (b) relations transitively joining to the former, and (c) a subset of their attributes that should be present in the result, according to the preferences registered for the user that poses the query. (For more details, we refer the interested reader to [20].) After its creation, the schema of the logical database subset is stored in the graph database as a pattern associated with the group that the user submitting the query belongs to.

Logical Subset Generator. A précis pattern selected from the previous step is enriched with tuples extracted from the database according to the query keywords, in order to produce the logical database subset. For this purpose, the Logical Subset Generator starts from the initial relations where tokens in Q appear. Then, more tuples from other relations are retrieved by (foreign-key) join queries starting from the initial relations and transitively expanding on the database schema graph following edges of the pattern. Joins on a précis pattern are executed in order of decreasing weight. In other words, a précis pattern comprises a kind of a "plan" for collecting tuples matching the query and others related to them. At the end of this phase, the logical database subset has been produced.

Translator. The Translator is responsible for rendering a logical database subset to a more user-friendly synthesis of results. This is performed by a semi-automatic method that uses templates over the database schema. In the context of this work, the presentation of a query answer is defined as a proper structured management of individual results, according to certain rules and templates predefined by a designer. The result is a user-friendly response through the composition of simple clauses.

In this framework, in order to describe the semantics of a relation R along with its attributes in natural language, we consider that relation R has a conceptual meaning captured by its name, and a physical meaning represented by the value of at least one of its attributes that characterizes tuples of this relation. We name this attribute the *heading* attribute and we depict it as a hachured rounded rectangle. For example, in Fig. 1, the relation ARTIST conceptually represents "artists" in real world; indeed, its name, ARTIST, captures its conceptual meaning. Moreover, the main characteristic of an "artist" is its name, thus, the relation ARTIST should have the NAME as its heading attribute. By definition, the edge that connects a heading attribute with the respective relation has a weight 1 and it is always present in the result of a précis query. A domain expert makes the selection of heading attributes.

The synthesis of query results follows the database schema and the correlation of relations through primary and foreign keys. Additionally, it is enriched by alphanumeric expressions called *template labels* mapped to the database graph edges.

A *template label*, label(u,z) is assigned to each edge $e(u,z) \in \mathbf{E}$ of the database schema graph $\mathbf{G}(\mathbf{V},\mathbf{E})$. This label is used for the interpretation of the relationship between the values of nodes u and z in natural language.

Each projection edge $e \in \mathbf{\Pi}$ that connects an attribute node with its container relation node, has a label that signifies the relationship between this attribute and the heading attribute of the respective relation; e.g., the BIRTH_DATE of an ARTIST (.NAME). If a projection edge is between a relation node and its heading attribute, then the respective label reflects the relationship of this attribute with the conceptual meaning of the relation; e.g., the NAME of an ARTIST. Each join edge $e \in \mathbf{J}$ between two relations has a label that signifies the relationship between the heading attributes of the relations involved; e.g., the WORK (.TITLE) of an ARTIST (.NAME). The label of a join edge that involves a relation without a heading attribute signifies the relationship between the previous and subsequent relations.

We define as the label l of a node `n` the name of the node and we denote it as l(`n`). For example, the label of the attribute node `NAME` is “name”. The name of a node is determined by the designer/administrator of the database. The template label `label(u,z)` of an edge `e(u,z)` formally comprises the following parts: (a) `lid`, a unique identifier for the label in the database graph; (b) l(`u`), the name of the starting node; (c) l(`z`), the name of the ending node; (d) $\texttt{expr}_1$, $\texttt{expr}_2$, $\texttt{expr}_3$ alphanumeric expressions. A simple template label has the form:

$$\texttt{label(u,z)} = \texttt{expr}_1 + l(\texttt{u}) + \texttt{expr}_2 + l(\texttt{z}) + \texttt{expr}_3$$

where the operator “+” acts as a concatenation operator.

In order to use template labels or to register new ones, we use a simple language for templates that supports variables, loops, functions, and macros.

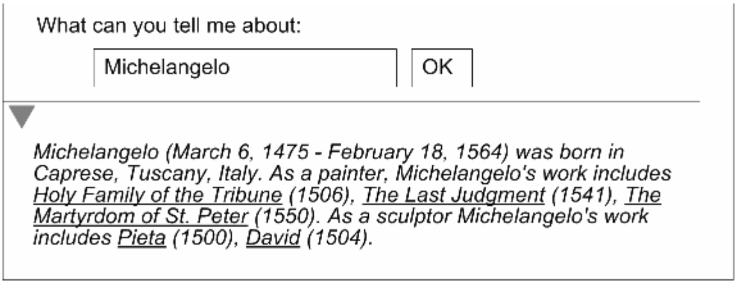
The translation is realized separately for every occurrence of a token. At the end, the précis query lists all the clauses produced. For each occurrence of a token, the analysis of the query result graph starts from the relation that contains the input token. The labels of the projection edges that participate in the query result graph are evaluated first. The label of the heading attribute comprises the first part of the sentence. After having constructed the clause for the relation that contains the input token, we compose additional clauses that combine information from more than one relation by using foreign key relationships. Each of these clauses has as subject the heading attribute of the relation that has the primary key. The procedure ends when the traversal of the databases graph is complete. For further details, we refer the interested reader to [21].

User Interface. The user interface of our prototype comprises a simple form where the user can enter one or more keywords describing the topic of interest. Currently, the system considers that query keywords are connected with the logical operator `or`. This means that the user can ask about “Michelangelo” `or` “Leonardo da Vinci”, but cannot submit a query about “Michelangelo” `and` Leonardo da Vinci”, which essentially would ask about the connection between these two entities/people.

Before using the system, a user identifies oneself as belonging to one of the existing groups, i.e. art reviewers or fans. Fig. 4 displays an example of a user query and the answer returned by the system. Underlined topics are hyperlinks. Clicking such a hyperlink, the user implicitly submits a new query regarding the underlined topic. For example, clicking on “*David*” will generate a new précis regarding this sculpture. Hyperlinks are defined on heading attributes of relations.

Although extensive testing of the system with a large number of users has not taken place yet, a small number of people have used the system to search for pre-selected topics as well as topics of their interest and reported their experience. This has indicated the following:

- The précis query answering paradigm allows users with little or no knowledge of the application domain schema, to quickly and easily gain an understanding of the information space.
- Naïve users find précis answers to be user-friendly and feel encouraged to use the system more.



What can you tell me about: Michelangelo

Michelangelo (March 6, 1475 - February 18, 1564) was born in Caprese, Tuscany, Italy. As a painter, Michelangelo's work includes Holy Family of the Tribune (1506), The Last Judgment (1541), The Martyrdom of St. Peter (1550). As a sculptor Michelangelo's work includes Pieta (1500), David (1504).

Fig. 4. Example précis query

- By providing précis of information as answers and hyperlinks inside these answers, the system encourages users to get involved in a continuous search-and-learn process.

5 Conclusions and Future Work

We have described the design, prototyping and evaluation of a précis query answering system with the following characteristics: (a) support of a keyword-based search interface for accessing the contents of the underlying collection, (b) generation of a logical subset of the database that answers the query, which contains not only items directly related to the query selections but also items implicitly related to them in various ways, (c) personalization of the logical subset generated and hence the précis returned according to the needs and preferences of the user as a member of a group of users, and (d) translation of the structured output of a précis query into a synthesis of results. The output is an English presentation of short factual information précis. As far as future work is concerned, we are interested in implementing a module for learning précis patterns based on logs of queries that domain users have issued in the past. In a similar line of research, we would like to allow users to provide feedback regarding the answers they receive. Then, user feedback will be used to modify précis patterns. Another challenge will be the extension of the translator to cover answers to more complex queries. Finally, we are working towards the further optimization of various modules of the system.

References

1. S. Agrawal, S. Chaudhuri, and G. Das. DBXplorer: A system for keyword-based search over relational databases. In *ICDE*, pp. 5-16, 2002.
2. I. Androutsopoulos, G.D. Ritchie, and P. Thanisch. Natural Language Interfaces to Databases - An Introduction. *NL Eng.*, 1(1), pp. 29-81, 1995.
3. G. Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti, and S. Sudarshan. Keyword searching and browsing in databases using BANKS. In *ICDE*, pp. 431-440, 2002.
4. A. Dusterhoft, and B. Thalheim. Linguistic based search facilities in snowflake-like database schemes. *DKE*, 48, pp. 177-198, 2004.

5. D. Florescu, D. Kossmann, and I. Manolescu. Integrating keyword search into XML query processing. *Computer Networks*, 33(1-6), 2000.
6. L. Guo, F. Shao, C. Botev, and J. Shanmugasundaram. XRank: Ranked keyword search over XML documents. In *SIGMOD*, pp. 16-27, 2003.
7. L. R. Harris. User-Oriented Data Base Query with the ROBOT Natural Language Query System. VLDB 1977: 303-312.
8. V. Hristidis, L. Gravano, and Y. Papakonstantinou. Effcient IR-style keyword search over relational databases. In *VLDB*, pp. 850-861, 2003.
9. V. Hristidis, Y. Papakonstantinou, and A. Balmin. Keyword proximity search on XML graphs. In *ICDE*, pp. 367-378, 2003.
10. IBM. *DB2 Text Information Extender. url: www.ibm.com/software/data/db2/extender/ textinformation/.*
11. G. Koutrika, A. Simitsis, and Y. Ioannidis. Précis: The essence of a query answer. In *ICDE*, 2006.
12. G. Marchionini. Interfaces for End-User Information Seeking. J. of the American Society for Inf. Sci., 43(2), 156-163, 1992.
13. U. Masermann, and G. Vossen. Design and implementation of a novel approach to keyword searching in relational databases. In *ADBIS-DASFAA*, pp. 171-184, 2000.
14. E. Metais, J. Meunier, and G. Levreau. Database Schema Design: A Perspective from Natural Language Techniques to Validation and View Integration. In *ER*, pp. 190-205, 2003.
15. E. Metais. Enhancing information systems management with natural language processing techniques. *DKE*, 41, pp. 247-272, 2002.
16. Microsoft. *SQL Server 2000. url: http://msdn.microsoft.com/library/.*
17. M. Minock. A Phrasal Approach to Natural Language Interfaces over Databases. In *NLDB*, pp. 181-191, 2005.
18. A. Motro. Constructing queries from tokens. In *SIGMOD*, pp. 120-131, 1986.
19. Oracle. *Oracle 9i Text. url: www.oracle.com/technology/products/text/.*
20. A. Simitsis, and G. Koutrika. Pattern-Based Query Answering. In *PaRMa*, 2006.
21. A. Simitsis, and G. Koutrika. Comprehensible Answers to Précis Queries. In *CAiSE*, pp. 142-156, 2006.
22. E. Sneiders. Automated Question Answering Using Question Templates That Cover the Conceptual Model of the Database. In *NLDB*, pp. 235-239, 2002.
23. V.C. Storey, R.C. Goldstein, H. Ullrich. Naive Semantics to Support Automated Database Design. *IEEE TKDE*, 14(1), pp. 1-12, 2002.
24. V.C. Storey. Understanding and Representing Relationship Semantics in Database Design. In *NLDB*, pp. 79-90, 2001.
25. A. Toral, E. Noguera, F. Llopis, and R. Munoz. Improving Question Answering Using Named Entity Recognition. In *NLDB*, pp. 181-191, 2005.
26. Q. Wang, C. Nass, and J. Hu. Natural Language Query vs. Keyword Search: Effects of Task Complexity on Search Performance, Participant Perceptions, and Preferences. In *INTERACT*, pp. 106-116, 2005.

The TIP/Greenstone Bridge: A Service for Mobile Location-Based Access to Digital Libraries

Annika Hinze, Xin Gao, and David Bainbridge

University of Waikato, New Zealand
{a.hinze, xg10, d.bainbridge}@cs.waikato.ac.nz

Abstract. *This paper introduces the first combination of a mobile tourist guide with a digital library. Location-based search allows for access to a rich set of materials with cross references between different digital library collections and the tourist information system. The paper introduces the system's design and implementation; it also gives details about the user interface and interactions, and derives a general set of requirements through a discussion of related work.*

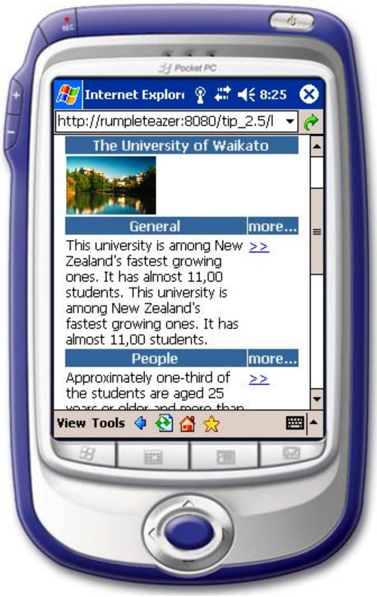
Keywords: *Digital Libraries, Tourist information, mobile system, location-based.*

1 Introduction

Digital Libraries provide valuable information for many aspects of people's lives that are often connected to certain locations. Examples are maps, newspaper articles, detailed information about sights and places all over the world. Typically, whenever people are at the location in question, computers are not close by to access the abundant information. In contrast, mobile tourist information systems give access to well-formatted data regarding certain sights, or information about travel routes. The abundance of information in history books or art catalogues has so far been (to all intents and purposes) excluded from these kind of systems.

This paper introduces the first known combination of a mobile tourist guide with a digital library. Location-based search allows access to a set of rich materials with cross references between different digital library collections and the tourist information system. The intention of the hybrid system is that a user, traveling with an internet connected mobile devise such as a pocketPC, is actively presented information based on their location (automatically detected through GPS) and a user profile that records their interests (such as architectural follies); when a passing detail seems particularly pertinent or piques their interest (hopefully the norm rather than the exception, otherwise the information the tourism system is providing is not useful) the user seamlessly taps into the "deeper" resources managed by the digital library that can better satisfy their quest for both more details and related information. Usage scenarios for location-based access

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 99–110, 2006.



This university is among New >> Zealand's fastest growing ones. It has almost 11,00 students. This university is among New Zealand's fastest growing ones. It has almost 11,00 students.

Approximately one-third of >> the students are aged 25

Fig. 1. TIP: location-based personalised information delivery

to digital library collections include looking up the pictures of van Gogh while being at their locations of origin in France, comparing old postcard photographs with current buildings, reading Goethe while traveling Italy.

Due to their open source nature and the authors' familiarity with the software TIP [6] and Greenstone [13] are used as the foundations of the hybrid system. The paper commences with a brief review of these two project, emphasising aspects pertinent to the work at hand. Next we discuss the challenges of a location-based bridge between the two systems (Section 2). We subsequently show the TIP/Greenstone service in use (Section 3). In Sections 4 and 5, we give details of the service's design and architecture, respectively. Related work is discussed in Section 6. Our paper concludes with a summary and an outlook to future research.

2 Background

This section describes the foundations of the TIP project and the Greenstone project that are necessary for this paper.

2.1 TIP Core

The Tourist Information Provider (TIP) System delivers location-based information to mobile users. The information delivered is based on a user's context, such as their current location, their interest in particular semantic groups of sights and topics, and their travel history. Semantic groups and topics are captured in a user's profile. Examples for groups of sights are public art, buildings, or beaches, topics may be history or architecture. The travel history of a user includes the locations/sights that the user visited and the user's feedback about these sights.

Figure 1 shows the TIP standard interface in a mobile emulator. The user is at the University of Waikato. Their profile is *groups*={*buildings; parks*}; *topics* = {*architecture; history*}. The university is displayed as a building close to the user's current position. In addition to the core functionality, TIP supports several services such as recommendations and travel navigation on maps (for details see [6]). The TIP system combines an event-based infrastructure and location-based service for dynamic information delivery. The heart of the system is a filter engine cooperating with a location engine. The filter engine selects the appropriate information from the different source databases based on the user and sight context. Changes in the user's location are transmitted to the TIP server, where they are treated as events that have to be filtered. For the filtering, the sight context and the user context are taken into account. The location engine provides geo-spatial functions, such as geo-coding, reverse geo-coding, and proximity search. For places that are currently of interest, the system delivers sight-related information.

2.2 Greenstone

Greenstone is a versatile open source digital library toolkit [13]. Countless digital libraries have been formed with it since its release on SourceForge in 2000: from historic newspapers to books on humanitarian aid; from eclectic multimedia content on pop-artists to curated first editions of works by Chopin; from scientific intuitional repositories to personal collections of photos and other document formats. All manner of topics are covered—the black abolitionist movement, bridge construction, flora and fauna, the history of the Indian working class, medical artwork, and shipping statistics are just a random selection. All manner of document formats are covered, including: HTML, PDF, Word, PowerPoint, and Excel; MARC, Refer, Dublin Core, LOM (Learning Object Metadata) and BibTeX metadata formats; as well as a variety of image, audio, and video formats. It also supports numerous standards including OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting), Z39.50 and METS to assist interoperability. See *www.greenstone.org* for more details.

For the pattern of use needed by this project Greenstone is an ideal vehicle, providing a rapid way to "tera-form" a haphazard mass of digital resources into an organized, manageable collection accessible through browsing structures as well as direct access through fielded and full-text searching. The end result is shaped by a configuration process. Of particular relevance here to the hybrid system is the algorithmic role metadata plays in shaping this process through its storage, access and ultimately presentation. Later we will see this aspect of Greenstone exploited—seeded through a gazetteer—for identifying and marking up place names.

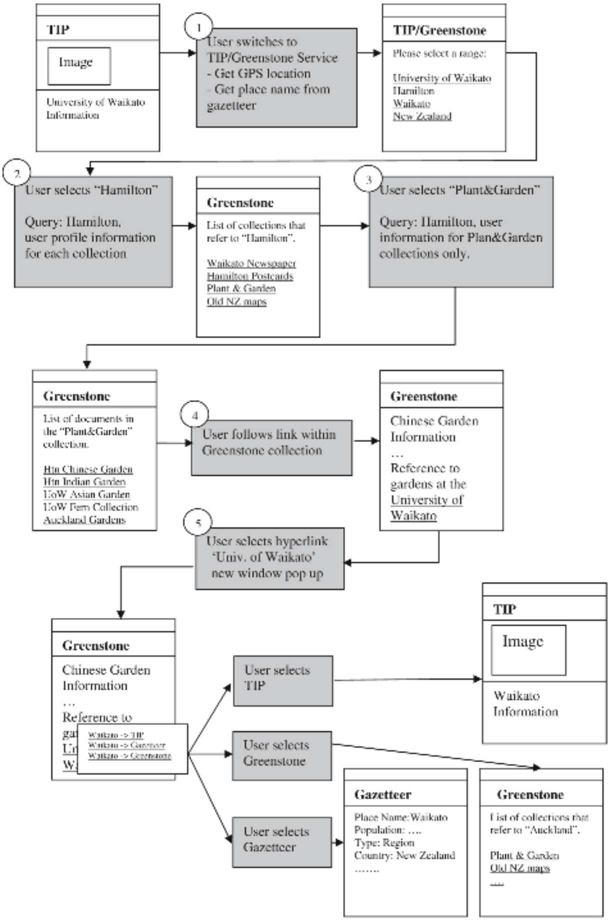


Fig. 2. Overview of example interaction with TIP/Greenstone

In addition to the production-level version of the software used for the cited examples above (known as Greenstone 2), for exploratory purposes a digital library framework that utilises web services is available for research-based work. Called Greenstone 3 and backwards compatible, this forms the digital library code base used for the hybrid system. In particular it allows for the fine-grained interaction necessary for the required functionality, as evidenced through the worked example in the next section.

3 Usage Scenarios and Interface

Here we demonstrate usage in a typical interactions scenario with the TIP/Greenstone Service: the user travels to a location using the TIP systems and then accesses more detailed information about this location from the digital library. An overview of the interactions for the example Gardens in Hamilton is given in Figure 2.

For this usage scenario, we follow a user Kate who visits the Waikato area in New Zealand. She uses TIP with the TIP/Greenstone Service when she arrives in the Waikato, near the university. Her initial view is that shown in Figure 1. (For clarity, we will show further TIP screens using screenshots taken from a browser window in Figure 3.) Kate then decides to look up the digital library collections that refer to her current location. When she switches to the page of TIP/Greenstone Service, the system will display regions and places that are near-by that she might want to search for in the collection repository provided by Greenstone. This is necessary as her location is the University of Waikato, close to the river Waikato, in the region Waikato, in New Zealand, on the north island, etc. All these locations could be used to search the library and the user can guide the selection. This step is referred to in the overview as Step 1 (see Figure 2). Based on Kate's selection, the system triggers a location-based search in DL collections. The user is presented with a list of all collections that refer to the selected region. (We will give the details about the system internal design for this step later).

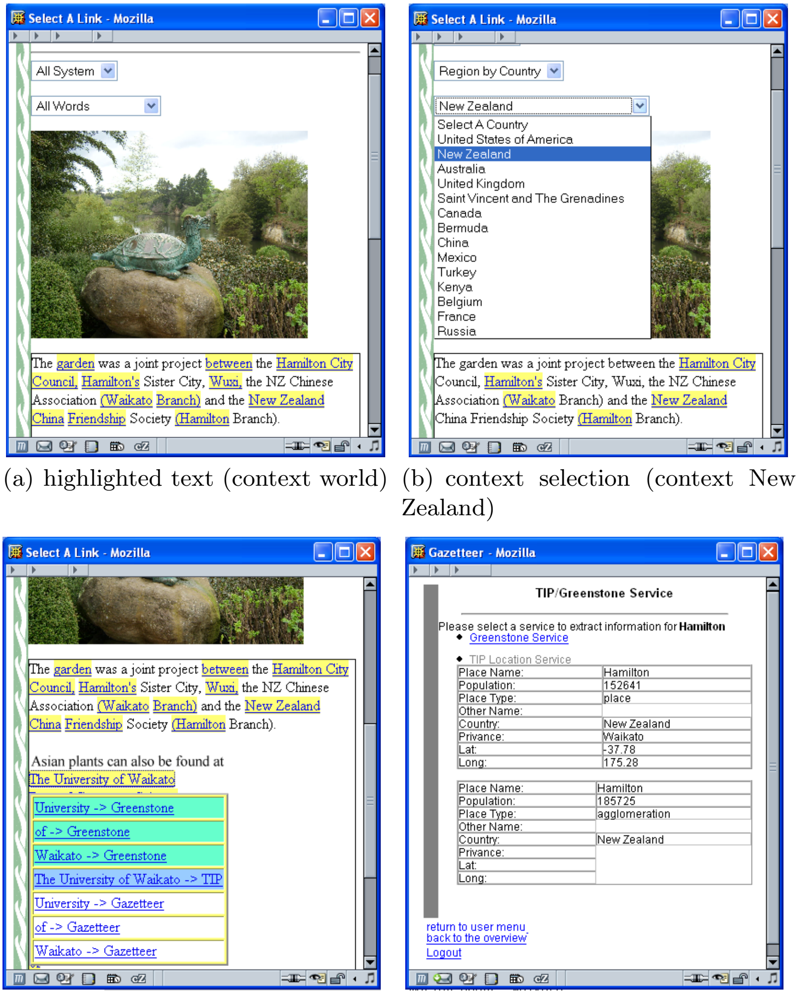
After selecting the region 'Hamilton' in Step 2, Kate has the choice between the *Hamilton postcard collection*, the *Waikato newspaper collection* and the *Plant&Garden collection*; she selects the Plant & Garden collection (Step 3). Amongst others, this collection contains references to the local Hamilton Gardens. Within the collection, Kate selects a document about the Chinese Garden in Hamilton (Step 5). Figure 3(a) shows the Greenstone interface with the Plant & Garden collection. Kate chose to indicate all place names in the document: a special feature of the TIP/Greenstone service. All words that are recognised as place names are highlighted. Kate can direct the focus for these highlights to particular countries (see Figure 3(b)).

Kate can now decide to lookup the highlighted places in TIP or in Greenstone. This link to TIP or to other programs is reached via a pull-down menu shown in Figure 3(c). The menu displays links only to existing data pages/documents. Different background colors indicate different target programs. One of the options is to display information about the places from the geo-gazetteer (a geographical database of place names worldwide). The gazetteer provides information about location, population, province, country. It displays this information in the location-context of the document, i.e., only locations within the selected country are displayed. Figure 3(d) shows the information about 'Hamilton' when selecting 'New Zealand'—showing only 2 of the 26 Hamiltons worldwide in the geo-gazetteer, the second referring to the wider conurbation that strays outside the designated city boundary.

4 Architecture

The TIP/Greenstone service connects TIP's service communication layer (similar to other services) to Greenstone as a third party application of Greenstone. Figure 4 shows the position of the service between TIP and Greenstone.

The communication with TIP is currently handled via TCP/IP and HTTP. Greenstone provides an interface of communication via SOAP for communication



The garden was a joint project between the Hamilton City Council, Hamilton's Sister City, Wuxi, the NZ Chinese Association (Waikato Branch) and the New Zealand China Friendship Society (Hamilton Branch).

(a) highlighted text (context world)

The garden was a joint project between the Hamilton City Council, Hamilton's Sister City, Wuxi, the NZ Chinese Association (Waikato Branch) and the New Zealand China Friendship Society (Hamilton Branch).

(b) context selection (context New Zealand)

The garden was a joint project between the Hamilton City Council, Hamilton's Sister City, Wuxi, the NZ Chinese Association (Waikato Branch) and the New Zealand China Friendship Society (Hamilton Branch).

Asian plants can also be found at The University of Waikato

(c) back link pop-up (context world)

Please select a service to extract information for **Hamilton**

- Greenstone Service
- TIP Location Service

Place Name:	Hamilton
Population:	152641
Place Type:	place
Other Name:	
Country:	New Zealand
Privance:	Waikato
Lat:	-37.78
Long:	175.28

Place Name:	Hamilton
Population:	185725
Place Type:	agglomeration
Other Name:	
Country:	New Zealand
Privance:	
Lat:	
Long:	

(d) gazetteer (context New Zealand)

Fig. 3. Overview of example interaction with TIP/Greenstone

to take place between its agents as well as communicate with third party programs; consequently the TIP/Greenstone service uses this protocol to connect to the message router of Greenstone, thereby giving it access to the full topology of a particular instantiation of a collection server. To initiate a fielded search based on location, for example, the user's profile information and current location are translated into XML format to call the Text-Query Service of the Greenstone collections. Search results are also in XML format, and so translated into HTML and for presentation to the user.

For the interaction with users, Greenstone uses HTML pages by default to present the response of activating the underlying services. A user's interaction with a web page will initiate a data transfer to the Library Servlet. The

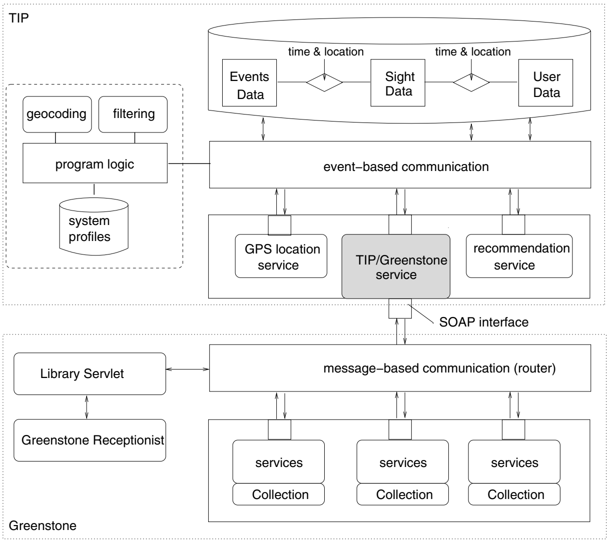


Fig. 4. Architecture of TIP and Greenstone with the TIP/Greenstone Bridge

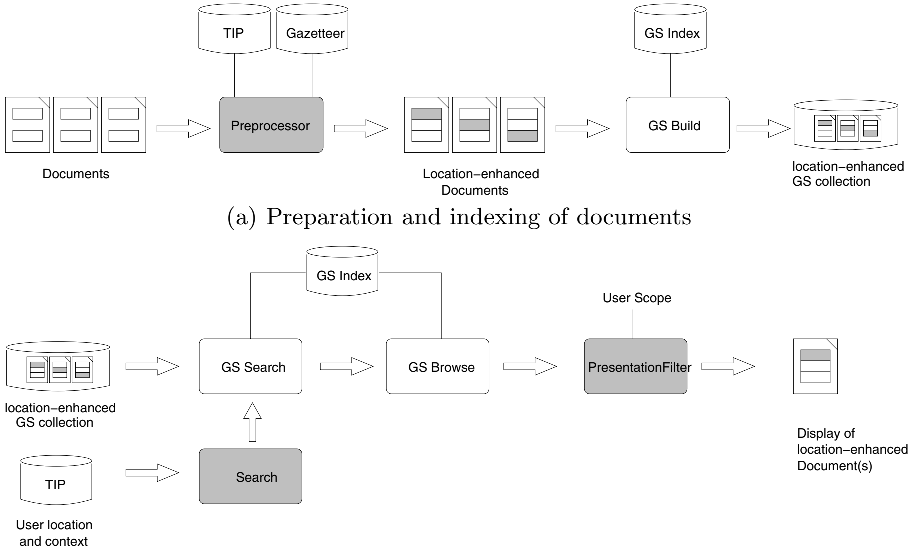
Library Servlet translates the information into XML format and forwards it to the Receptionist. The Receptionist is an agent responsible for the knowing which channels of communication to use to satisfy a request and, upon the return of XML-borne messages, how to merge and present the information to the user. The TIP/Greenstone Service performs similar interactions between users and Greenstone, except it does not work through the receptionist. Instead its contact point, as mentioned above, is the message router, through which it can mesh with the array of services on offer. Effectively the TIP/Greenstone Service takes on the role of receptionist (in Greenstone terminology)—factoring in information retrieved in a user's TIP profile and current location—deciding which channels of communication to use and how to merge and present the resulting information.

5 Detailed Design

Handling location-based documents in a mobile setting falls into two phases: preparation of documents and retrieval. The steps are described in detail in the subsequent paragraphs. Figure 5 gives an overview of the phases.

Preprocessing: Location identification
The pre-processing phase locates place names in the documents of a collection and enhances the documents. To recognize those place names that contain more than one word in the gazetteer and TIP system, a place name window has been designed. Figure 6 shows an example of how a place name window works.

The documents are first analysed for their textual context—all HTML markup is ignored. All remaining words are analysed using the place name window and


TIP
Gazetteer
GS Index
Documents
Preprocessor
Location−enhanced Documents
GS Build
location−enhanced GS collection
GS Index
User Scope
location−enhanced GS collection
GS Search
GS Browse
PresentationFilter
TIP
Search
User location and context
Display of location−enhanced Document(s)

(a) Preparation and indexing of documents

(b) Location-based retrieval and display

Fig. 5. Two phases of location-based handling of DL documents

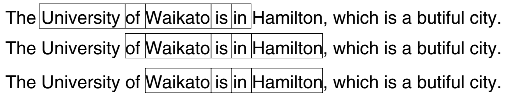

The University of Waikato is in Hamilton, which is a butiful city.
The University of Waikato is in Hamilton, which is a butiful city.
The University of Waikato is in Hamilton, which is a butiful city.

Fig. 6. A sliding place name window to identify composite and nested place names

the place name validator of the gazetteer and TIP. The current array of words in the current place name window defines the set of words to be analysed. The initial size of the place name window can be set by the user; it is currently set to five words. By changing the size of the sliding window, place names that are nested within longer place names (‘Waikato’ within ‘University of Waikato’) are also recognised. The validator returns the name of the country in the gazetteer and/or the site in TIP.

The preprocessor marks up the documents with location-based information (in the form of Java Script) that provides the multiway hyperlink seen in Figure 3(c). The first parameter includes all the place names found in the longest place name. For instance, if the place name is New York, then the parameter will be York and New York. The second parameter refers to the countries. The third parameter is a reference for the place name in the TIP system. The next one is the ID for the gazetteer.

DL Collections for Location-based access
Collections that are used by the TIP/Greenstone Service, are stored in the digital library using the standard storage facility of Greenstone. Currently, the collections in the DL need to be built in a particular way to interoperate with TIP.

More specifically, existing collections have to be pre-processed before the can enter the standard build process for collections. The preprocessor assigns location information to each occurrence of a place name in the collection's documents. Location information contains details on related countries or coordinates. This data is encoded within the collection's HTML documents using Java Script. Currently, only collections with HTML documents are supported in the TIP/Greenstone bridge. After the collections have been built by Greenstone they will be kept as normal Greenstone collections. An improvement would be to integrate the preprocessing step into the main building code, but this was not done due to time constrains. By delaying the application of place name metadata markup to the point where the internal document model (text and structure) has been built, then the technique would equally apply to the wide range of textual formats handled by Greenstone cited in Section 2.

Location-based search
Special features have been implemented for location-based search and location highlighting. Location information is obtained from a geo-gazetteer stored in the TIP database. The information for the gazetteer has been imported from *www.world-gazetteer.com*. The TIP/Greenstone Service has access to the postgreSQL database used in TIP in addition to the information in the gazetteer, to store information about locations. The TIP database is used in this service to load the information about the place rather than using the information in the gazetteer. The coordinates of places are queries in the TIP database to calculate the nearest place in the gazetteer. TIP uses the PostGIS spatial extension which supports spatial querying.

All accessed documents and collections from the digital library are filtered according to the location-restriction. Additional restrictions on countries help identify appropriate place names as a large number of English words can also be identified as place names (e.g., 'by', which is a village in France). In addition and on request, stop words are excluded from the location filter. Finally, the filter introduces hyperlinks into the document that allow for references to TIP, the gazetteer, and Greenstone. Original links are preserved.

Location-based presentation
The Presentation-Filter component uses the location markup (Java Script in the documents) to highlight places that are within the current scope of the user. To determine the current user scope, the user is offered a drop-down box of all countries that appear in the text. The user can then select the current scope. Place names are highlighted and hyperlinks are added that link to related pages in TIP, Greenstone, and the Gazetteer. This list of hyperlinks with different target services are accessible on left-click in a pop-up menu using dynamic HTML to effectively give web pages multiway hyperlinks. References to other services can be easily added. Original hyperlinks in the document need to be treated with care: they are still visually indicated on the page as hyperlinks. In addition, the list of hyperlinks contains the phrase 'original link' which then refers to the original target of the document's hyperlink. If a nested place name (a name within another name) is not within the current location scope, the longer name

is highlighted but the shorter (nested) one is removed from the hyperlink list. For further implementation details and a first-cut performance study see [3].

6 Discussion of Related Work

To the best of our knowledge, no combination of a digital library with a mobile tourist system has been developed previously. Mobile tourist information systems focus mainly on providing (short) structured information and recommendations. Examples systems are AccessSights [7], CATIS [10], CRUMPET [11], Cyber-Guide [1], Guide [2], Gulliver's Genie [9]. This lack of ability for location-based access of rich sources was our motivation for combining the tourist information system with a digital library resource. It is also the case that most research in the area of electronic guides has focused on indoor users.

In considering the digital library aspect to the work and how it relates to other digital library projects, we identify the following key requirements for the hybrid system:

1. bi-directional interoperability
2. geographical "aware" digital library
3. generic collection building
4. extensible presentation manipulation
5. markup restructuring
6. fine-grained interaction
7. fielded searching

Bi-directional interoperability allows the two sub-systems to be able to work in unison, and some shared notion of geographical information is needed in which to have a meaningful exchange. To populate the digital library resource it is necessary to have a digital library system with a workflow that includes flexible indexing and browsing controls (generic collection building), to allow the tera-forming of initially formless source documents. To provide the representative functionality shown in the worked example (Section 3), control at the presentation-level is required that is context based and includes fine-grained interaction with sub-systems to garner the necessary information and the filtering and restructuring of markup.

The related issue of spatial searches for place names and locations has been addressed in digital libraries (for example, in the Alexandria digital library project [4, 12]). Access to geo-spatial data is typically given in Geo-Information Systems (GIS), which use spatial database for storage. A spatial digital library with a GIS viewing tool has also been proposed [5]. In these systems, the focus lies on the geo-spatial features and their presentation on maps or in structured form. Rich documents with ingrained location information are not the focus. Nor do these project have the ability to be installed by third party developers and populated with their own content.

Contemporary digital library architectures such as Fedora [8] that include a protocol as part of their design satisfy the requirement for fine-grained interaction, and given that the *de facto* for digital libraries is presentation in a web browser there is a variety of ways manipulation of presentation through restructuring markup etc. can be achieved, for instance the dynamic HTML approach deployed in our solution. Fedora, however, only digests one canonical format of document, so does not meet the generic building requirement without an extra preprocessing step. Like the preprocessing step we added to Greenstone to augment it with place name metadata, this would be straightforward to do; indeed, the same metadata place name enhancement would be required also.

7 Conclusions

In conclusion, this paper has described a bridge between a mobile tourist information system and a digital library. This bridge service allows for location-based access of documents in the digital library. We explored the usage of the proposed hybrid system through a worked example. We gave an overview of the architecture and details of the implementation design.

Our system is an example of the trend through which digital libraries are integrated into co-operating information systems. Moreover, this work represents a focused case-study of how digital library systems are being applied to our increasingly mobile IT environments, and our experiences with the project encourage us to pursue further research towards interoperability between TIP and Greenstone. Although the work is centered on these two systems, through an analysis of general requirements we have outlined the key attributes necessary for developing co-operative hybrid information systems that combine mobile location information with digital libraries.

References

1. G. Abowd, C. Atkeson, J. Hong, S. Long, R. Kooper, and M. Pinkerton. Cyberguide: A mobile context-aware tour guide. *ACM Wireless Networks*, 3:421–433, 1997.
2. K. Cheverst, K. Mitchell, and N. Davies. The role of adaptive hypermedia in a context-aware tourist guide. *Communications of the ACM*, 45(5):47–51, 2002.
3. X. Gao, A. Hinze, and D. Bainbridge. Design and implementation of Greenstone service in a mobile tourist information system. Technical Report X/2006, University of Waikato, March 2006.
4. M. Goodchild. The Alexandria digital library project. *D-Lib Magazine*, 10, 2004.
5. P. Hartnett and M. Bertolotto. Gisviewer: A web-based geo-spatial digital library. In *Proceedings of the 5th International Workshop on Database and Expert Systems Applications (DEXA 2004), 30 August - 3 September 2004, Zaragoza, Spain*, pages 856–860, 2004.
6. A. Hinze and G. Buchanan. The challenge of creating cooperating mobile services: Experiences and lessons learned. In V. Estivill-Castro and G. Dobbie, editors, *Twenty-Ninth Australasian Computer Science Conference (ACSC 2006)*, volume 48 of *CRPIT*, pages 207–215, Hobart, Australia, 2006. ACS.
7. P. Klante, J. Krsche, and S. Boll. AccesSights – a multimodal location-aware mobile tourist information system. In *Proceedings of the 9th Int. Conf. on Computers Helping People with Special Needs (ICCHP'2004)*, Paris, France, July 2004.

8. C. Lagoze, S. Payette, E. Shin, and C. Wilper. Fedora: An architecture for complex objects and their relationships. *Journal of Digital Libraries*, 2005. Special Issue on Complex Objects.
9. G. O'Hare and M. O'Grady. Gulliver's genie: A multi-agent system for ubiquitous and intelligent content delivery. *Computer Communications*, 26(11):1177–1187, 2003.
10. A. Pashtan, R. Blattler, and A. Heusser. Catis: A context-aware tourist information system. In *Proceedings of the 4th International Workshop of Mobile Computing*, Rostock, Germany, 2003.
11. S. Poslad, H. Laamanen, R. Malaka, A. Nick, P. Buckle, and A. Zipf. CRUMPET: Creation of user-friendly mobile services personalised for tourism. In *Proc. 3G2001 Mobile Communication Technologies*, London, U.K., Mar. 2001.
12. T. R. Smith, G. Janee, J. Frew, and A. Coleman. The Alexandria digital earth prototype. In *ACM/IEEE Joint Conference on Digital Libraries*, pages 118–119, 2001.
13. I. H. Witten and D. Bainbridge. *How to Build a Digital Library.* Elsevier Science Inc., 2002.

Towards Next Generation CiteSeer: A Flexible Architecture for Digital Library Deployment

I.G. Councill[1], C.L. Giles[1]
E. Di Iorio[2], M. Gori[2], M. Maggini[2], and A. Pucci[2]

[1] School of Information Sciences and Technology, The Pennsylvania State University,
332 IST Building University Park, PA 16802
{icouncil, giles}@ist.psu.edu
[2] Dipartimento di Ingegneria dell'Informazione, University of Siena,
Via Roma, 56. Siena, Italy
{diiorio, marco, maggini, augusto}@dii.unisi.it

Abstract. CiteSeer began as the first search engine for scientific literature to incorporate Autonomous Citation Indexing, and has since grown to be a well-used, open archive for computer and information science publications, currently indexing over 730,000 academic documents. However, CiteSeer currently faces significant challenges that must be overcome in order to improve the quality of the service and guarantee that CiteSeer will continue to be a valuable, up-to-date resource well into the foreseeable future. This paper describes a new architectural framework for CiteSeer system deployment, named CiteSeer Plus. The new framework supports distributed indexing and storage for load balancing and fault-tolerance as well as modular service deployment to increase system flexibility and reduce maintenance costs. In order to facilitate novel approaches to information extraction, a blackboard framework is built into the architecture.

1 Introduction

The World Wide Web has become a staple resource for locating and publishing scientific information. Several specialized search engines have been developed to increase access to scientific literature including publisher portals such as the ACM Portal[1] and IEEE Xplore[2] as well as other academic and commercial sites including the Google Scholar[3]. A key feature common to advanced scientific search applications is citation indexing [3]. Many popular commercial search services rely on manual information extraction in order to build citation indexes; however, the labor involved is costly. Autonomous citation indexing (ACI) [4] has emerged as an alternative to manual data extraction and has proven to

[1] http://portal.acm.org/portal.cfm
[2] http://ieeexplore.ieee.org
[3] http://scholar.google.com

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 111–122, 2006.

be successful despite some loss of data accuracy. Additionally, the ACI model has traditionally been coupled with autonomous or semi-autonomous content acquisition. In this approach, focused crawlers are developed to harvest the web for specific types of documents, in this case academic research documents, in order to organize distributed web content within a single repository. Automatic content acquisition is particularly useful for organizing literature that would otherwise be difficult to locate via general search engines [8].

CiteSeer [4] emerged as one of the first focused search engines to freely provide academic papers, technical reports, and pre-prints, and is also the first example of a working ACI system. CiteSeer consists of three basic components: a focused crawler or harvester, the document archive and specialized index, and the query interface. The focused spider or harvester crawls the web for relevant documents in PDF and PostScript formats. After filtering crawled documents for academic documents, these are then indexed using autonomous citation indexing, which automatically links references in research articles to facilitate navigation and evaluation. Automatic extraction of the context of citations allows researchers to determine the contributions of a given research article quickly and easily; and several advanced methods are employed to locate related research based on citations, text, and usage information. Additional document metadata is extracted from each document including titles, author lists, abstracts and reference lists, as well as the more recent addition of author information such as affiliations and contact information [6] as well as acknowledgement information [5]. CiteSeer is a full text search engine with an interface that permits search by document or by numbers of citations or fielded searching, not currently possible on general-purpose web search engines.

CiteSeer has proven its usefulness to the computer and information science communities. The CiteSeer installation at Penn State University[4] currently receives over one million requests and serves over 25 GB of information daily. The CiteSeer service is currently being made more available to the world community through the advent of several mirrors. At the time of this writing there are CiteSeer mirrors hosted at MIT, Switzerland, Canada, England, Italy, and Singapore in various stages of completion. However, CiteSeer currently faces significant challenges of interoperability and scalability that must be overcome in order to improve the quality of the services provided and to guarantee that CiteSeer will continue to be a valuable, up-to-date resource well into the foreseeable future.

The current architecture of the CiteSeer application is monolithic, making system maintenance and extension costly. Internal system components are not based on any established standards, such that all interoperability features incorporated have necessarily been crafted as wrappers to exposed functionality. The resulting lack of integration reduces the potential of CiteSeer to serve the research community. Additionally, as the CiteSeer collection grows (to over 730,000 documents as of the time of this writing), query latencies are rising and document updates are becoming increasingly cumbersome as the system pushes the boundaries of its current architecture.

[4] http://citeseer.ist.psu.edu

Recently, other ACI-enabled search engines for scientific literature have been developed, including Google Scholar. Although Google Scholar indexes at least an order of magnitude more documents than CiteSeer, CiteSeer remains competitive as an open archive and offers more features. A separate effort that has shown much promise is OverCite, a re-implemenation of CiteSeer within a peer-to-peer architecture based on distributed hash tables [15].

In this paper we present our own re-invention of CiteSeer, currently named CiteSeer Plus. This work builds on a previous architectural proposal for digital libraries [13]. CiteSeer Plus is based upon a new architecture designed to be flexible, modular, and scalable. As CiteSeer is currently operated within an academic environment with a focus on research as well as production, we have developed a framework that allows scalable, distributed search and storage while easing deployment of novel and improved algorithms for information extraction as well as entirely new service features.

The resulting architecture is oriented toward a collection of deployed services instead of a traditional web search engine approach. Each service component can be treated as a stand-alone application or as part of a larger service context. Users and programs can interact directly with individual services or with the entire system through web-based service front-ends such as a traditional search engine interface, consistent with ideas emerging from *Web 2.0* [11].

2 Project Goals

Flexibility. CiteSeer's current monolithic architecture limits the extensibility of the system. Information extraction routines are difficult to upgrade or change since they are tightly coupled with other system components. Not only does this cause maintenance difficulty, but it also limits the potential scope of the CiteSeer system. Adopting a highly modular service-oriented architecture will make the system more easily extendable with new services and more adaptable to different content domains. This is a core requirement for a next-generation CiteSeer service. Although an API has been developed for the existing CiteSeer [13], the API does not expose internal system functionality that is needed for a powerful extension environment. To alleviate this problem, each service module should carry its own API. This will allow service extensions to combine components in a flexible manner without incurring the overhead of refactoring existing code, and will allow the system to be more easily extensible to novel content domains.

Performance. A next-generation CiteSeer system must show improvements over the current system in terms of both query processing and update performance. Due to the current indexing and database framework, CiteSeer shows significant performance degradation when handling more than five simultaneous queries. Traffic spikes often account for more than 30 simultaneous queries and as many as 130 simultaneous connections have been observed. The resulting performance drop often limits the query response times to well below acceptable standards, in many cases turning users away outright. The new system should be able to handle at least 30 simultaneous queries without significant performance degradation. In addition, CiteSeer currently indexes no more than 3-4 papers per minute, resulting in poor speed for acquiring new content. The update processes are large batch operations that typically take three days for every two weeks of content acquisition. To improve the freshness of information in the repository, it is desirable for a next-generation CiteSeer architecture to handle content updates quickly in an iterative process, so new content can be made available immediately after acquisition.

Distributed Operation. Although CiteSeer is currently implemented as a collection of processes that interoperate over network sockets, the architecture does not currently support redundant service deployment. This situation is mitigated through the use of Linux Virtual Server for service load balancing and fail-over; however, this increases maintenance demands and does not support distributed operation in a WAN environment. There is no support for propagating updates to mirrors without large file copies containing much redundant information. The new system should be natively capable of distributed operation with no single point of failure and should be easily extendable to support incremental updates over a WAN deployment.

3 System Features and Architecture

This section details the features supported by the CiteSeer Plus framework as well as its architecture. CiteSeer Plus is designed to be a flexible platform for digital library development and deployment, supporting standard digital library features as well as plugins for advanced automation. In keeping with the goals presented in Section 2, the feature set is expandable based on application or domain requirements and the user interface to the application is arbitrary, to be built on top of a rich system API. An experimental prototype of a CiteSeer Plus deployment is publicly available[5].

The CiteSeer Plus system architecture is highly modular. In the following sections every module is presented and module interactions are discussed. The system architecture is organized in four logical levels as shown in Figure 1.

The *Source Level* contains document files and associated data. The *Core Level* contains the central part of the system in which document and query processing occurs. The *Interface Level* offers interface functions to allow the communication between the Core Level and services that can be developed using CiteSeer Plus (in the Service Level). This level is implemented as a collection of Web Services. Finally, the *Service Level* contains every service that is running on top of the CiteSeer Plus system.

Figure 2 maps the levels to the actual system architecture. At the Core Level are the sets of master and slave indexing nodes. These sets contain redundant indexing nodes tailored for specific tasks within the CiteSeer Plus system, and are the fundamental processing nodes. A single node is made of different subcomponents. Figure 3 shows the details of a master indexing node. We can describe these nodes by following a typical paper lifecycle through an indexing node.

[5] http://p2p.science.unitn.it/cse

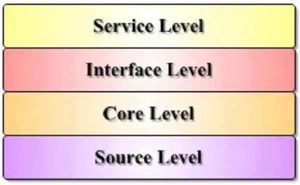


Fig. 1. Logical levels

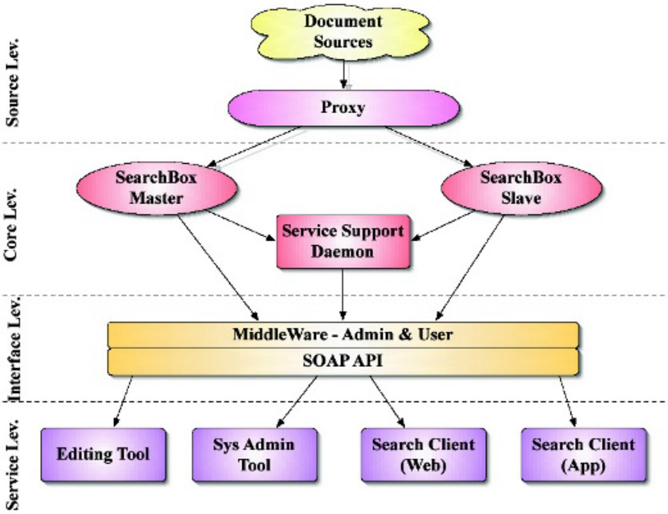


Fig. 2. System architecture overview

The system is agnostic regarding the method of content acquisition. New content may be harvested by a crawler, received from an external library, or submitted by users, so long as documents are posted to the system via a supported acquisition interface. Once a paper has been received it is stored in the *PDF cache* to guarantee persistence of a document in the original format, then submitted to a document processing workflow for integration into the system data. The paper encounters a *PDF parser* whose duty is to extract text from the original file and produce a new XML-based representation. This new document contains text and some layout information such as text alignment, size, style, etc. Next the raw XML file enters the metadata extraction subsystem. This subsystem is composed of several modules, including a *BlackBoard Engine* that is used to run a pool of experts (shown as *EXP 1*, *EXP 2*, . . . , *EXP N* in Figure 3) that cooperate to extract information from the document. This process is presented in more detail in Section 5. This process outputs an XML document that contains all tagged metadata.

Finally the paper is ready to be indexed: the labeled XML is stored in the *XML cache* (to make it available for later retrieval) and passed to the *indexer*.

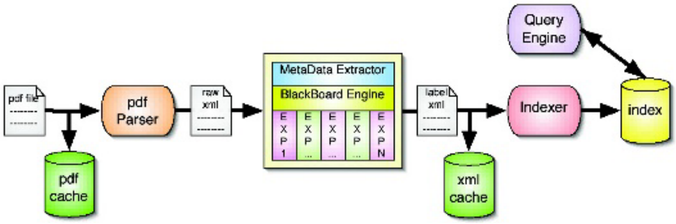


Fig. 3. Indexing node detailed structure

At this point the *Query Engine* will be able to provide responses to user or system queries involving the indexed document. Metadata elements are stored in separate index fields, enabling complex queries to be built according to various document elements. Every indexing node is able to communicate with the other system components by exposing a set of methods as a web service. The entire indexing process takes place in on-line mode, such that a paper entering the system will enter one or more indexing nodes for immediate consumption by the system.

In addition to normal indexing nodes (called *master nodes*) there are also *slave nodes*. Slave nodes are a lighter version of master nodes; their inner structure is just the same as seen in Figure 3, with the exception that slave nodes do not maintain any kind of cache (no PDF cache nor XML cache). Furthermore, their indexes contain only metadata slices (such as title, author, abstract and citation slices), but they do not contain generic text slices, which support full-text queries. Both master and slave nodes can be deployed redundantly for load balancing. During initial indexing, a paper can be routed to any number of slave nodes but must be routed to at least one master node, in order to allow the system to provide full-text indexing and caching. Slave nodes are provided in order to support frequent operations such as citation browsing, graph building, and paper header extraction (a header contains just title, author, abstract and references) since those operations do not require access to a full-text index. In this way, performance can be improved by adding new slave nodes that do not incur large additional storage requirements. Slave nodes can also be used to support system processing for graph analysis and the generation of statistics without affecting performance for user queries; however, only a single master node is needed to run a CiteSeer Plus system.

It is also possible to split the indexes among different machines (in this case the controller will send a query to all of them and then organize the different responses received). At the same time, indexes can be redundant; that is, the same indexes can be managed by different mirror nodes running on different computers in order to improve system performance through load balancing. In Figure 4 we show a typical system configuration.

In this deployment we have divided the index into two parts (A and B), so every time a document is accepted by the system, the controller decides which subindex will receive the document, such that indexes are balanced. Nodes in

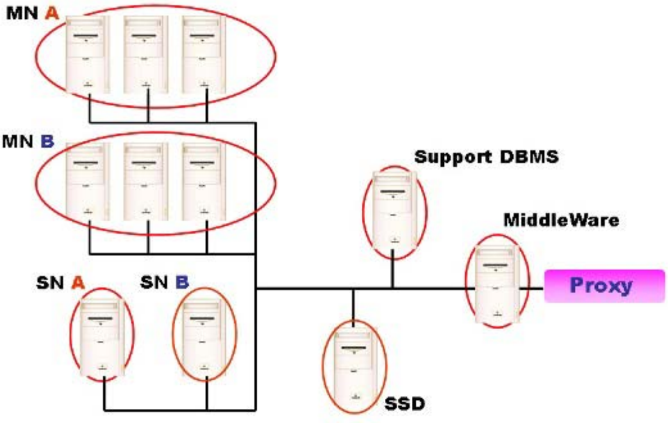


Fig. 4. Example of system deployment

the same node set have the same indexes to support index redundancy. In this example "MN A" (master node set of subindex A) contains three computers running three separated and identical master node instances, and "SN A" provides support to "MN A" nodes. In this case "SN A" contains only one slave node, but, in general, it can be a set of slave nodes. The same configuration is kept for the "B" (in this case we have "MN B" and "SN B"). In this scenario, if a user submits a full-text query the controller will route the query to a master node chosen from the "MN A" set and one from "MN B", so the system, in this sample configuration, is able to provide service for up to three concurrent users just the same as one by sharing the workload among redundant master node mirrors inside "MN A" and "MN B". The same situation happens when a query does not involve a full-text search, but is just referred to metadata indexes. The only difference in this case is the fact that slave nodes ("SN A" and "SN B") will respond to the query instead of master nodes.

At the Interface Level we find the *Middleware*, which is the active part of the external SOAP API. This component converts API methods into procedure calls on the services provided by the components in the Core Level. The Middleware contains methods to perform user authentication control in order to determine whether a system user is authorized to perform the requested operations. The Middleware also manages the controller threads and performs query and paper routing in order to maintain consistency in the distributed and redundant sets of Master and Slave Nodes. Every operation regarding resource distribution and redundancy is performed in this module.

Each system component exposes public methods through the *SOAP API*, allowing the development of discrete services using the CiteSeer Plus framework. The Service Level uses the API to define prescribed usage scenarios for the system and interfaces for user control. This level contains HTML forms and representations for user and administrative interaction. Some exemplar services that have been built include tools to add or remove documents and correct document metadata, deployment configuration tools, and search interfaces for users (a web application) or programs (via SOAP).

4 Citation Graph Management

A document citation graph is a directed graph where the nodes correspond to the documents and edges correspond to citation relationships among the documents. A document citation graph is useful for deriving bibliometric analyses such as computing document authorities and author importance as well as to perform social network analysis. In order to construct a document citation graph all citations contained in each document must be identified and parsed, and then the citations must be matched to corresponding document records. CiteSeer Plus uses an approach that differs in many ways from the legacy CiteSeer.

CiteSeer's method could be defined as a *"hard approach"*. Each citation is parsed using heuristics to extract fields such as title, authors, year of publication, page numbers and the citation identifier (used to mark the citation in the body text). The fields of each citation are compared with one another based on a string distance threshold in order to cluster citations into groups representing a single document. Finally, the metadata from each citation group is compared to existing document records in order to match the citations to documents. Citations to a given paper may have widely varying formats; hence, developing rules for citation field identification can be very time consuming and error prone. CiteSeer's approach relies heavily on off-line computations in order to build the document citation graph. If no document is found to match a citation group, all citations in the group are *unsolved*, and cannot be solved until the next graph update, even if a matching document enters the system beforehand.

The CiteSeer Plus approach could be defined as a *soft approach*. Our method is less computationally costly and can be performed online, in an approach similar to the SFX system [16]. The process of building the citation graph in CiteSeer Plus is query-based; that is, the citations are solved using queries performed in the *query module*. The Indexer allows metadata to be stored in different sub-indexes (*slices*) and so a query can be performed on a specific slice of the main index. Subfields parsed from citations are used to perform complex document queries on appropriate index slices and the top document is found to match a citation if it's similarity to the query surpasses a given threshold. In the other direction, to find citations matching a new document, CiteSeer Plus makes a query using all the words of the document title and authors. This query is performed on the citation slice; thus the query results are all documents that have a citation containing some words of the query.

Master nodes do not cache the document citation graph since they have to provide query results that are as fresh as possible. However, slave nodes can use a query result caching mechanism in order to improve performance at the cost of reduced information freshness. Repository statistics are built using slave nodes, but user queries operate on the master node. When a user tries to follow a citation, this produces a corresponding query on the master node and the user will obtain one or more documents that are likely to match the citation. This framework relieves workload on dynamic components that handle user queries while allowing detailed statistics and graph management activities to be handled online within separate components.

5 Metadata Extraction System

Metadata extraction is the most difficult task performed by an automated digital library system for research papers. In the literature, there are two main approaches to information extraction: knowledge engineering and machine learning. In the knowledge engineering approach, the extraction rules used by the system are constructed manually by using knowledge about the application domain. The skill of the knowledge engineer plays a large role in the level of system performance, but the best performing systems are often handcrafted. However, the development process can be very laborious and sometimes the required expertise may not be available. Additionally, handcrafted rules are typically brittle and do not perform well when faced with variation in the data or new content domains. CiteSeer uses this approach, employing information about the computer science document styles (or templates) to extract metadata.

In the machine learning approach, less human expertise regarding template styles is required when customizing the system for a new domain. Instead, someone with sufficient knowledge of the domain and the task manually labels a set of training documents and the labeled data is used to train a machine learning algorithm. This approach is more flexible than the knowledge engineering approach, but requires that a sufficient volume of training data is available. In the last decade, many techniques have been developed for metadata extraction from research papers. There are two major sets of machine learning techniques in the metadata extraction literature. Generative models such as Hidden Markov Models (HMM) (e.g. [14], [9]) learn a predictive model over labeled input sequences. Standard HMM models have difficulty modeling multiple non-independent features of the observation sequence, but more recently Conditional Random Fields (CRF) have been developed to relax independence assumptions [7]. The second set of techniques is based on discriminative classifiers such as Support Vector Machines (SVM) (e.g. [6]). SVM classifiers can handle large sets of non-independent features. For the sequence labeling problem, [6] work in a two stage process: first classifying each text line independently in order to assign it a label, then adjusting these labels based on an additional classifier that examines larger windows of labels. The best performance in metadata extraction from research papers has been reached by McCallum and Peng in [12] using CRFs. The CiteSeer Plus metadata extraction system has been built to maximize flexibility such that it is simple to add new extraction rules or extraction models into the document processing workflow. In our metadata extraction system, different kinds of models can be used which have been trained for different or the same extraction tasks using various techniques, including but not limited to HMM, CRF, regular expression, and SVM classifiers. The CiteSeer Plus metadata extraction system is based on a blackboard architecture ([10], [1], [2]) such that extraction modules can be designed as standalone processes or within groups of modules with dependencies. A blackboard system consists of three main components:

Knowledge Sources (in our framework these are named Experts): independent modules that specialize in some part of the problem solving. These experts can be widely different in their inference techniques and in their knowledge representation.

BlackBoard: a global database containing the input data, partial solutions and many informational items produced by experts to support the problem solving.

Control component: a workflow controller that makes runtime decisions about the course of problem solving. In our framework, the control component consists of a set of special experts called *scheduling experts* that are able to schedule the knowledge sources registered in the framework. The scheduling expert is chosen by the controller components based on the problem solving strategy that is employed and the kinds of metadata that the system needs to progress. Using different scheduling experts, it is possible to change the problem solving strategy dynamically in order to experiment with various learning strategies.

Although an individual expert can be independent from all the other experts registered in the framework, each expert can declare its information dependences, that is, all the information that it needs to work. The *control component* activates the expert when all these dependences are satisfied. As such, experts can be activated when all the information required by the expert has been extracted and stored on the *BlackBoard module.* The experts declare their skills (the information they can extract) to the *Control component*, such that during the problem solving (metadata extraction), at the right moment the control component can activate the experts, and the controller can reason about which intermediary experts must be employed in order to reach a later result. The BlackBoard groups similar information and registers expert accuracies based on the *prior expertise*[6] declared by each expert. In this way, if more than one expert produces the same (or similar) kinds of information, the accuracy value of that information will be computed as the joint confidence among the experts.

An example configuration may group experts into three classes or *functional levels*, although the framework does not restrict the processing workflow. The first level is the *Entity Recognition* level. In this level are all the experts able to give words a specific semantic augmentation, including part-of-speech tagging and recognition of named entities such as first or last name, city, country, abbreviation, organization, etc. Experts at this level will be activated first for processing workflows. The second level is the *Row Labeling* level. At this level are all the experts able to classify a paper line with one or more defined labels such as author, title, affiliation, citation, section title and so on. The experts at this level classify the paper lines using a document representation supplied by the *Document module*, a framework object able to elaborate the document

[6] The prior expertise is a measure of expert ability (F score) on a standard dataset.

structure by supplying a representation based on many different features regarding line contents, layout and font styles. Row labeling can be an iterative process, reclassifying lines based on tagged context in subsequent passes. The last level is the *Metadata Construction* level. Using all the extracted information from the previous levels, the experts at this level can build the final metadata record for a document.

6 Summary

This paper has presented a new version of the CiteSeer system, showing significant design improvements over its predecessor. The new system reproduces every core feature of the previous version within a modular architecture that is easily expandable, configurable, and extensible to new content domains. Increased flexibility is obtained through a design based on customizable plug-in components (for the metadata extraction phase) and the extensive use of web service technology to provide an interface into every system component. CiteSeer Plus can also be a useful tool for researchers or other developers interested in information retrieval and information extraction, as CiteSeer Plus can be used as a powerful yet easy to use framework to test new ideas and technologies by developing third party applications that bind with specific components of the CiteSeer Plus framework.

Acknowledgments

We thank Nicola Baldini and Michele Bini (FocuSeek.com) for fruitful discussions, suggestions and support during the system design and development process. We also thank Fausto Giunchiglia and Maurizio Marchese (University of Trento, Italy) for fruitful discussions that have aided the evolution of the system.

References

1. B. L. Buteau. A generic framework for distributed, cooperating blackboard systems. *Proceedings of the 1990 ACM annual conference on Cooperation*, p.358-365, February 20-22, 1990.
2. H. Chen , V. Dhar. A knowledge-based approach to the design of document-based retrieval systems. *ACM SIGOIS Bulletin*, v.11 n.2-3, p.281-290, Apr. 1990.
3. E. Garfield. Science Citation Index - A new dimension in indexing. *Science*, 144, pp. 649-654, 1964.
4. C.L. Giles, K. Bollacker and S. Lawrence. *CiteSeer: An Automatic Citation Indexing System*, Digital Libraries 98: Third ACM Conf. on Digital Libraries, ACM Press. New York, 1998, pp. 89-98.
5. C.L. Giles and I.G. Councill. Who gets acknowledged: measuring scientific contributions through automatic acknowledgement indexing. *PNAS*, 101, Number 51, pp. 17599-17604, 2004.

6. H. Han, C. Lee Giles, E. Manavoglu, H. Zha, Z. Zhang, E. A. Fox. Automatic Document Metadata Extraction using Support Vector Machines. *Proceedings of the 2003 Joint Conference on Digital Libraries (JCDL03)*, 2003.
7. J. Lafferty, A. McCallum, and F. Pereira. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. *In International Conference on Machine Learning*, 2001.
8. S. Lawrence, C. Lee Giles. Searching the World Wide Web. *Science*, 280, Number 5360, pp. 98-100, 1998.
9. T. R. Leek. Information extraction using hidden Markov models. *Masters thesis, UC San Diego*, 1997.
10. H. Penny Nii. Blackboard systems: The blackboard model of problem solving and the evolution of blackboard architectures. *The AI Magazine*, VII(2):38–53, Summer 1986.
11. T. O'Reilly. What Is Web 2.0 Design Patterns and Business Models for the Next Generation of Software. *http://www.oreillynet.com/pub/a/oreilly/tim/news /2005/09/30/what-is-web-20.html*
12. F. Peng and A. McCallum. Accurate information extraction from research papers using conditional random fields. *Proceedings of Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics(HLT-NAACL)*, pages 329336 (2004).
13. Y. Petinot, C. Lee Giles, V. Bhatnagar, P. B. Teregowda, H. Han, I. Councill. A Service-Oriented Architecture for Digital Libraries. *ICSOC04*, November 15-19, 2004.
14. K. Seymore, A. McCallum and R. Rosenfeld. Learning hidden Markov model structure for information extraction. *In Papers from the AAAI-99 Workshop on Machine Learning for Information Extration*, pages 3742, July 1999.
15. J. Stribling, I.G. Councill, M.F. Kaashoek, R. Morris, and S. Shenker. Overcite: A cooperative digital research library. In *Proceedings of The International Workshop on Peer-To-Peer Systems (IPTPS 05)*, Ithaca, NY, 2005 .
16. H. Van de Sompel, P. Hochstenbach. Reference linking in a hybrid library environment. Part 1: Frameworks for linking. *D-Lib Magazine*, v.5 n.4, 1999.

Digital Object Prototypes: An Effective Realization of Digital Object Types

Kostas Saidis[1], George Pyrounakis[2], Mara Nikolaidou[2], and Alex Delis[1]

[1] Dept. of Informatics and Telecommunications
[2] Libraries Computer Center
University of Athens
University Campus, Athens, 157 84, Greece
{saiko, forky, mara, ad}@di.uoa.gr

Abstract. Digital Object Prototypes (DOPs) provide the DL designer with the ability to model diverse types of digital objects in a uniform manner while offering digital object type conformance; objects conform to the designer's type definitions automatically. In this paper, we outline how DOPs effectively capture and express digital object typing information and finally assist in the development of unified web-based DL services such as adaptive cataloguing, batch digital object ingestion and automatic digital content conversions. In contrast, conventional DL services require custom implementations for each different type of material.

1 Introduction

Several formats and standards, including METS [10], MPEG-21 [15], FOXML [7] and RDF [11] are in general able to encode heterogeneous content. What they all have in common is their ability to store and retrieve arbitrary specializations of a digital object's constituent components, namely, files, metadata, behaviors and relationships [9]. The derived digital object typing information – that is, which components constitute each different type of object and how each object behaves – is not realized in a manner suitable for effective use by higher level DL application logic including DL modules and services [13].

Our main objective is to enhance how we express and use the types of digital objects independently of their low-level encoding format used for storage. Digital object prototypes (DOPs) [13] provide a mechanism that uniformly resolves digital object typing issues in an automated manner. The latter releases DL users such as cataloguers, developers and designers from dealing with the underlying complexity of typing manually. A DOP is a *digital object type definition* that provides a detailed specification of its constituent parts and behaviors. Digital objects are conceived as instances of their respective prototypes. DOPs enable the generation of *user-defined types* of digital objects, allowing the DL designer to model the specialities of each type of object in a fine-grained manner, while offering an implementation that guarantees that all objects conform to their type automatically. Using DOPs, the addition of a new digital object type requires no custom development and services can be developed to operate directly on all types of material without additional coding for handling "special" cases.

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 123–134, 2006.

DOPs assist in dealing with important "every day" DL development issues in a unified way: how to speed up and simplify cataloguing, how to automate content ingestion, how to develop effective web interfaces for presenting and manipulating heterogeneous types of digital objects. In this paper, we focus on the benefits offered by the deployment of DOPs in the development of high level services in Pergamos, the University of Athens DL. In particular, we point out how web based services such as browsing, cataloguing, batch ingestion and automatic digital content conversion cope with any type of DOP defined object, while having all services reside in a single, uniform implementation.

The remainder of the paper is organized as follows. Section 2 provides a detailed description of the current implementation of DOPs and pinpoints how DOPs assist on the development of uniform yet effective DL services. In Section 3 we present several DOP examples originating from Pergamos collections. Finally, Section 4 concludes the paper discussing related and future work.

2 Digital Object Prototypes in Pergamos

We have implemented DOPs in Java. As depicted in Figure 1a, DOPs operate atop the repository / storage layer of the DL (in Pergamos we use FEDORA [14]).

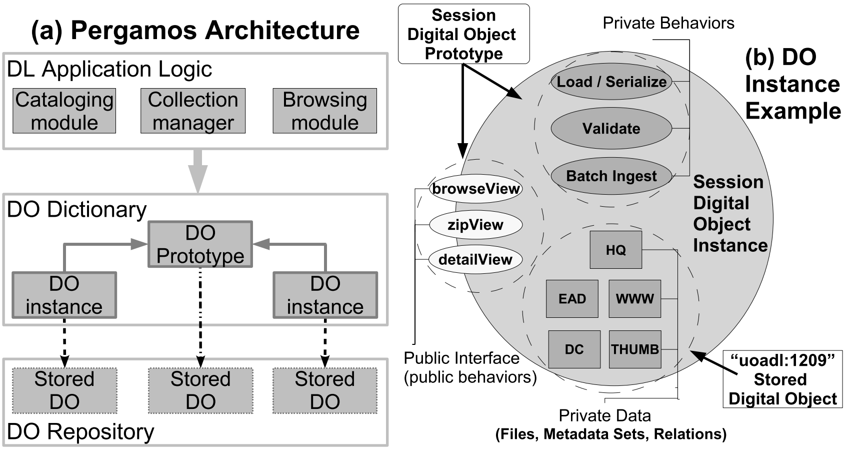


Fig. 1. (a) The 3-tier Pergamos architecture incorporating the "type enforcement" layer of DO Dictionary [13] and (b) A digital object instance as composed by its respective prototype and the underlying stored digital object

The *DO Dictionary* layer of Figure 1a exposes the DOPs API to high level DL services or the application logic. The underlying repository's "mechanics" remain hidden, since all service functionality is directed through DOPs. We define DOPs in terms of XML documents, that are loaded by the *DO Dictionary* at bootstrap time. These XML documents provide the type specification that is translated to a Java representation wrapped by the DOPs API. At runtime, the *DO Dictionary* loads stored digital objects from the repository and generates Java artifacts named digital object instances that conform to their respective DOP definition. High level services operate on digital object instances; any modification occurring in an instance's data is serialized back to the repository when the object is saved.

In order to illustrate how DOPs effectively realize digital object types, in this section we use examples drawn from the Senate Archive's Session Proceedings collection in Pergamos DL. We model Session Proceedings using `Session` and `Page` DOPs; each Senate `Session` is modelled as a complex object containing `Pages`. Figure 1b depicts the runtime representation of a `Session` digital object instance, while Figure 2 illustrates the definition of the `Session` DOP, encoded in XML. The `Session` instance reflects the specifications found in the `Session` DOP. The instance's behaviors are defined in the DOP the instance conforms to, while its metadata, files and relations are loaded from its underlying stored digital object.

```xml
<prototype id="Session">
<MDSets><!-- Metadata definition -->
 <MDSet id="dc">
  <label lang="en">Dublin Core Metadata</label>
  <datastream id="DC" MDType="descriptive"
     loader="gr.uoa.dl.core.xml.StandardLoader"
     serializer="gr.uoa.dl.core.xml.DCSerializer"/>
  <fields>
   <field id="dc:date" isMandatory="true"
     isRepeatable="false" isHidden="false"
     validation ="gr.uoa.dl.core.validation.DateFormat">
    <label lang="en">Date</label>
   </field>
   <field id="dc:identifier_physical" isMandatory="true"
     isRepeatable="false" isHidden="true">
    <label lang="en">Call number</label>
   </field>
   ...
  </fields>
 </MDSet>
 <MDSet id="ead">
  <label lang="en">EAD like Metadata</label>
  <datastream id="EAD" MDType="descriptive"
     loader="gr.uoa.dl.core.xml.StandardLoader"
     serializer="gr.uoa.dl.core.xml.EADSerializer"/>
  <fields>
   <field id="did_unitid"/>
   ...
  </fields>
 </MDSet>
 <mappings>
  <mapping id="identifier">
   <from="ead.did_unitid">
   <to="dc.dc:identifier_physical"/>
  </mapping>
 </mappings>
</MDSets>
```

```xml
<files><!-- Files definition -->
 <file id="zip" type="container" datastream="ZIP">
  <label lang="en">ZIP file</label>
  <mime-type id="application/zip"/>
  <batchIngest targetTypeId="page" targetFileId="hq"/>
 </file>
</files>
<relations><!-- Relationships definition -->
    <structure>
        <childType>Page</childType>
    </structure>
    <references allowCustomURL="false" allowCustomDO="true"/>
</relations>
<behaviours><!-- Behaviours definition -->
 <schemes>
  <scheme id="browseView" isDefault="true">
   <label lang="en">Short View</label>
   <element id="MDSets.dc.dc:identifier"/>
   <element id="MDSets.dc.dc:title"/>
   <element id="MDSets.dc.dc:date"/>
  </scheme>
  <scheme id="zipView">
   <label lang="en">Short View</label>
   <element id="MDSets.dc.dc:title"/>
   <element id="files.zip"/>
  </scheme>
  <scheme id="detailView">
   <label lang="en">Detail View</label>
   <element id="MDSets.dc.dc:identifier"/>
   <element id="MDSets.dc.dc:identifier_physical"/>
   <element id="MDSets.dc.dc:title"/>
   <element id="MDSets.dc.dc:date"/>
   ...
  </scheme>
 </schemes>
</behaviours>
</prototype>
```

Fig. 2. The `Session` prototype defined in XML terms

DOP definitions are encoded in XML as depicted by the `Session` DOP of Figure 2 and are made up of four parts according to [9]: (a) metadata element set definitions expressed in the `MDSets` XML section, (b) digital content specifications expressed in the `files` section, (c) relationships, defined in the `relations` section and (d) behaviors, defined in the `behaviors` XML section. In the following we provide a detailed description of each of these four definition parts, while, in parallel, we discuss how these type definitions are interpreted at runtime. It is worth pointing out that, although most of the examples we use herein originate from object input scenarios, the automatic type conformance offered by DOPs covers all aspects of digital object manipulation. The DOPs framework is not a static digital object model. On the contrary, it can be conceived as a framework that allows users to define their own digital object models.

2.1 Behaviors in DOPs

The behaviors of a digital object constitute the set of operations it supports. All the instances of the same DOP share the same behaviors; for example, all Session Proceedings behave in the same manner. This is reflected by the fact that with DOPs, behaviors are defined only in the object's respective prototype and are automatically bound to the digital object instance at runtime by the *DO Dictionary*.

DOPs implement digital object types by drawing on the notions of the OO paradigm. In order to support OO encapsulation, our approach distinguishes private from public behaviors. Private behaviors refer to operations that are executed by the digital object instance in a private fashion, hidden from third parties. For example, validations of metadata element values are private behaviors that are executed by instances according to their DOP specification, without user intervention. Private behaviors are triggered on specific events of the digital object instance's lifecycle; for instance, when a DL service updates the metadata of an object. Private behaviors are implicitly defined in the DOP, as described in the examples presented later in this section. On the other hand, public behaviors constitute the interface through which third parties can interact with the digital object instance at hand. Public behaviors are explicitly defined in a DOP and are described in Section 2.5.

2.2 Metadata Elements in DOPs

DOPs support the use of multiple metadata element sets for describing different digital object characteristics [9,10]. There are three ways to specify a metadata element set in a DOP: (a) as a standard element set, such as the Dublin Core (DC) [3], (b) as a user-defined extension of a standard element set (e.g. qualified DC) or (c) as a totally custom element set. In detail, a DOP specifies:

- the individual metadata sets contained in the objects of this type, supplied with an identifier and a multi-lingual label and description.
- the specific elements that constitute each metadata set. Each element is designated by an identifier, desired labels and descriptions, and additional behavioral characteristics expressed in terms of private behaviors.
- the possible mappings among elements of the various metadata sets.

As the `MDSets` section of Figure 2 illustrates, `Session` objects are characterized using a qualified DC metadata set, called `dc`. Due to the archival nature of the material, we also use a second, custom element set called `ead`, that follows the principles of Encoded Archival Description (EAD) [6], yet without encoding the EAD Finding Aid in its entirety.

In what follows, we describe the metadata handling capabilities of DOPs and provide appropriate examples drawn from the MDSets specifications found in the Session prototype of Figure 2.

Automatic loading & serialization of Metadata sets: Loading and serialization of metadata sets are private behaviors, both executed by the DOP behind the scenes. For example, if a DL service requests the dc metadata set values of a Session digital object instance, the DOP specified loader is used to load the corresponding element values from the underlying stored digital object. Respectively, whenever a DL service stores the digital object instance to the repository, the DOP supplied serializer is used to serialize each metadata set to the appropriate underlying format. Loaders and serializers are defined in the datastream XML section of the MDSet definition. Each DOP is allowed to define its custom loading / serialization plugins, given that they constitute valid implementations of the respective Loader and Serializer Java interfaces supplied by the *DO Dictionary*. The Session DOP, for example, uses the StandardLoader plugin to load the metadata of Session Proceedings objects.

Behavioral characteristics of Metadata elements: The DOPs metadata specification inherently offers additional behavioral characteristics for each metadata element. These characteristics are exploited by DL services on a case to case basis for each element. DOPs define behavioral characteristics in terms of XML attributes of the respective field definitions appearing in the MDSet specification. In DOPs, we support the following behavioral characteristics:

- isMandatory: the instance will throw an exception if the metadata element is to be saved with a null value.
- isHidden: advices the UI to hide the element from end-users.
- isRepeatable: the metadata element is allowed to have multiple values. The UI service adjusts accordingly, by supplying the cataloguer with the ability to insert multiple values or by displaying the values to the end-user in a list.
- validation: digital object instances apply the given validation whenever they are called to set values to the element. The validation occurs just before the user-supplied values are serialized and sent to the repository. DOPs support user-defined, pluggable validations, given that they implement the Validation interface provided by the *DO Dictionary*. For example, the definition of the dc:date element in Figure 2 specifies the use of a validation that checks whether respected values conform to the date format selected by the Senate Archive's cataloguing staff.

Mappings among Metadata Elements: The Session DOP of Figure 2 maps ead:unitid to dc:identifier_physical. A mapping between elements represents another example of a private behavior. Whenever the value of the ead:unitid element is modified, the digital object propagates its new value to the dc:identifier_physical. In Session objects, the mappings are created from selected ead elements to members of the dc metadata set. This is performed in order to allow us to offer cross-collection search to our users, given that FEDORA only supports DC metadata searches. With the use of DOP-based mappings we supply Pergamos with such search capabilities, without having to limit our material description requirements to DC metadata only or force our cataloguing staff to provide redundant information for both ead and dc metadata sets.

2.3 Digital Content in DOPs

With regard to digital content, a prototype:

- specifies the various files and their respective formats,
- provides the necessary information required for converting a primary file format to derivatives in order to automate and speed up the ingestion process,
- enables batch ingestion of content and automatic creation of the appropriate digital objects.

Listing 1.1 depicts the files configuration of the Senate Archive's Page DOP. The latter specifies that Page objects should contain three file formats, namely a high quality TIFF image (hq), a JPEG image of lower quality for web display (web) and a small JPEG thumbnail image for browsing (thumb). In what follows we describe batch ingestion and content conversion capabilities of DOPs.

```
<files>
 <file id="hq" type="primary" datastream="HQ">
  <label lang="en">High Quality Image</label>
  <mime-type id="image/tiff">
   <conversion target="web" task="convRes" hint="scale:0.6,quality:0.7"
      mimeType="image/jpeg" converter="gr.uoa.dl.core.conv.ImageConverter"/>
   <conversion target="thumb" task="convRes"
      hint="width:120,height:120,quality:0.6"
      mimeType="image/jpeg" converter="gr.uoa.dl.core.conv.ImageConverter"/>
  </mime>
 </file>
 <file id="web" type="derivative" datastream="WEB">
  <label lang="en">Web Image</label>
  <mime-type id="image/jpeg"/>
 </file>
 <file id="thumb" type="derivative" datastream="THUMB">
  <label lang="en">Thumbnail Image</label>
  <mime-type id="image/jpeg"/>
 </file>
</files>
```

Listing 1.1. The files section of the Page prototype

Automatic Digital Content Conversions: Each file format is characterized either as primary or derivative. In the case of files of Senate Archive's Page objects, as defined in the files section of Listing 1.1, the hq file is primary, referring to the original digitized material. The web and thumb files are treated as derivatives of the primary file, since the prototype's conversion behavior can generate them automatically from the hq file. Conversion details reside in the conversion section of each file specification. After the ingestion of the primary file, the digital object instance *executes the conversions residing in its prototype automatically.*

We support three conversion tasks, namely (a) `convert`, used to convert a file from one format to another, (b) `resize`, used to resize a file while maintaining its format and (c) `convRes`, used to perform both (a) and (b). Each task is carried out by the Java module supplied in the `converter` attribute, offering flexibility to users to provide their own custom converters. The converter is supplied with a `hint`, specifying either the required width and height of the resulting image in pixels, the scale factor as a number within (0, 1) or the `derivative`'s quality as a fraction of the original. In the case of `Page` objects (Listing 1.1), the `hq` file is converted to a `web` JPEG image using compression quality of 0.7 and resized using a scale factor of 0.6. Additionally, the `hq` file is also converted to a `thumb` JPEG image using compression quality 0.6 and dimensions equal to 120 x 120 pixels. The `Page` instance stores both derivatives in the FEDORA datastreams specified in the `datastream` attribute of their respective `file` XML element.

Batch Digital Object Ingestion: We also use DOPs to automate digital object ingestion. The `files` section of the `Session` prototype (Figure 2), depicts that `Session` objects are complex entities that contain no actual digital content but act as containers of `Page` objects. However, the `Session` prototype defines a `zip` file that is characterized as `container`. Containers correspond to the third supported file format. If the user uploads a file with the `application/zip` mime type in a `Session` instance, the latter initiates a `batchIngest` procedure. The `Session` DOP's `batchIngest` specification expects each file contained in the zip archive to abide to the `hq` file definitions of the `Page` prototype. In other words, if the user supplies a `Session` instance with a zip file containing TIFF images, as the `Session` `zip` file definition requires, the instance will automatically create the corresponding `Page` digital objects. Specifically, the `Session` `batchIngest` procedure extracts the zip file in a temporary location and iterates over the files it contains using the file name's sort order. If the file at hand abides to the `Page`'s `primary` file format:

a. Creates a new `Page` digital object instance.

b. Adds the `Page` instance to the current `Session` instance (as required from structural relationships described in Section 2.4).

c. Adds the file to the `Page` instance at hand. This will trigger the automatic file conversion process of the `Page` prototype, as outlined earlier.

Should we consider a `Session` comprised of 120 `Page` objects, then the ingestion automation task, supplied by DOPs, releases the user from creating 120 digital objects and making 240 file format conversions manually.

2.4 Relationships in DOPs

DOPs specify the different relationships that their instances may be allowed to participate in. Currently, DOPs support the following relationships:

- *Internal Relationships*: Digital objects reference other DL pertinent objects.

- *Structural Relationships*: These model the "parent / child" relationships generated between digital objects that act as containers and their respective "children".

- *External Relationships*: Digital object reference external entities, providing their respective URLs.

A `Session` object is allowed to contain `Page` objects; this specification appears in the `relations` section of the `Session` DOP (Figure 2). The existence of a `structure` specification in the `Session` prototype yields the following private behavior in the participating entities:

- Every `Session` object instance maintains a list of all the digital object identifiers the instance contains.
- Every `Page` instance uses the `dc:relation_isPartOf` element to hold the identifier of its parent `Session`.

Finally, the `references` part of the `relation` section informs DL services whether custom relationships are supported by this type of object. In the `Session` DOP of Figure 2, the `references` value guides UI services to allow the cataloguer to relate `Session` instances only with DL internal objects and not with external entities.

2.5 Public Behaviors in DOPs

We define public behaviors in DOPs using the notion of *behavioral scheme*. A behavioral scheme is a selection of the entities that are part of a digital object. Behavioral schemes are used to generate projections of the content of the digital object. Figure 2 illustrates the `behaviors` section of the `Session` prototype, which defines three behavioral schemes, namely `browseView`, `zipView`, and `detailView`. The `browseView` scheme supplies the user with a view of the digital object instance containing only three elements of the qualified DC metadata set, namely `dc:identifier`, `dc:title` and `dc:date`. Respectively, `zipView` generates a projection containing the `dc:title` metadata element and the `zip` file, while `detailView` provides a full-detail view of the object's metadata elements. This way, the DL designer is able to generate desired "subsets" of the encapsulated data of the digital object instance at hand for different purposes.

Execution of public behavior is performed by the invocation of a high level operation on a digital object instance, supplying the desired behavioral scheme. High level operations correspond to the actions supported by the DL modules. For example, the cataloguing module supports the `editObject`, `saveObject` and `deleteObject` actions, the browsing module supports the `browseObject` action, while object display module supports the `viewObject` action. At this stage, all Pergamos DL modules support only HTML actions:

- `viewObject("uoadl:1209", shortView)`: Dynamically generates HTML that displays the elements participating in the `shortView` of the "uoadl:1209" object in read-only mode. The *DO Dictionary* will first instantiate the object via its respective `Session` DOP (Fig. 1b). The new instance "knows" how to provide its `shortView` elements to the object display module.
- `editObject("uoadl:1209", zipView)`: Dynamically generates an HTML form that allows the user to modify the instance's elements that participate

in zipView. This view is used by the digitization staff in order to upload the original material and trigger the batch ingestion process, as described earlier in this section.

- editObject("uoadl:1209", detailView): Generates an HTML form that displays all the metadata elements of the given instance in an editable fashion. This is used by the cataloguing staff in order to edit digital object's metadata. The cataloguing module uses the behavioral characteristics described in Section 2.2 (e.g. isMandatory, isRepeatable) to generate the appropriate, type-specific representation of the digital object.

- saveObject("uoadl:1209", zipView): Saves "uoadl:1209" instance back to the repository. Only the zipView scheme elements are modified. Cataloguing module knows how to direct the submission of the web form generated by its aforementioned editObject action to saveObject. Respectively, cataloguing deleteObject action is bound to a suitable UI metaphor (e.g. a "delete" button of the web form). The scheme supplied to deleteObject is used to generate a "deletion confirmation view" of the digital object.

The execution of public behaviors is governed by the particular scheme at hand, while the DOP specifications enable DL application logic to adjust to the requirements of each element participating in the scheme.

3 Organization of Collections in Pergamos Using DOPs

Currently, Pergamos contains more than 50,000 digital objects originating from the Senate Archive, the Theatrical Collection, the Papyri Collection and the Folklore Collection. Table 1 provides a summary of the DOPs we generated for modeling the disparate digital object types of each collection, pinpointing the flexibility of our approach. It should be noted that DOPs are defined with a collection-pertinent scope [13] and are supplied with fully qualified identifiers, such as folklore.page and senate.page, avoiding name collisions. These identifiers apply to the object's parts, too; folklore.page.dc metadata set is different from the senate.page.dc set, both containing suitable qualifications of the DC element set for different types of objects.

a. Folklore Collection Folklore Collection consists of about 4,000 handwritten notebooks created by students of the School of Philosophy. We modeled the Folklore Collection using the Notebook, Chapter and Page DOPs. Notebooks are modeled as complex objects that reflect their hierarchical nature; the Notebook DOP allows notebooks to contain Chapter objects, which in turn are allowed to contain other Chapter objects or Page objects. Notebooks are supplied with metadata that describe the entire physical object, while Chapter metadata characterize the individual sections of the text. Finally, Page objects are not supplied with metadata but contain three files, resembling the definition of the Senate Archive's Pages provided in Listing 1.1.

b. Papyri Collection This collection is comprised of about 300 papyri of the Hellenic Papyrological Society. We modeled papyri using the Papyrus DOP, consisting of a suitable DC qualification and four file formats. The orig file format

Table 1. A summary of the DOPs we generated for four Pergamos collections

a. Folklore Collection

DOP	Metadata	Files	Relationships
Notebook	dc	none	contains Chapter or Page
Chapter	dc	none	contains Chapter or Page
Page	none	hq, web, thumb, hq to web, hq to thumb conversions	none

b. Papyri Collection

DOP	Metadata	Files	Relationships
Papyrus	dc	orig, hq, web, thumb, hq to web, hq to thumb conversions	none

c. Theatrical Collection

DOP	Metadata	Files	Relationships
Album	custom → dc	zip triggers batch import	contains Photo
Photo	niso → dc	hq, web, thumb, hq to web, hq to thumb conversions	none

d. Senate Archive's Session Proceedings

DOP	Metadata	Files	Relationships
Session	ead → dc	zip triggers batch import	contains Page
Page	none	hq, web, thumb, hq to web, hq to thumb conversions	none

corresponds to the original papyrus digitized image, while hq refers to a processed version, generated for advancing the original image's readability. The orig image is defined as primary, without conversions. The hq image, which is also defined as primary, is the one supplied with the suitable conversion specifications that generate the remaining two derivative formats, namely web and thumb.

c. Theatrical Collection Theatrical Collection consists of albums containing photographs taken from performances of the National Theater. Each Photo digital object contains three different forms of the photograph and is accompanied by the metadata required for describing the picture, either descriptive (dc) or technical (niso). As in the case of Senate Session Proceedings, mapping are used to to map niso elements to dc. Albums do not themselves contain any digital content, since they act as containers of Photo digital objects. However, Albums are accompanied by the required theatrical play metadata, encoded in terms of a custom metadata set, that is also mapped to dc.

d. Senate Archive The Senate Archive's Session Proceedings has been discussed in Section 2.

4 Discussion and Related Work

To our knowledge, DOPs provide the first concrete realization of digital object types and their enforcement. Our approach draws on the notions of the OO paradigm, due to its well established foundations and its well known concepts. Approaches on the formalization of OO semantics [2,12] show that the notion of objects in OO languages and the notion of digital objects in a DL system present significant similarities, yet in a different level of abstraction. [1] defines OO systems in terms of the following requirements:

- *encapsulation*: support data abstractions with an interface of named operations and hidden state,
- *type conformance*: objects should be associated to a type,
- *inheritance*: types may inherit attributes from super types.

At this stage, DOPs fulfill the encapsulation and type conformance requirements. The inclusion of inheritance is expected to provide explicit polymorphic capabilities to DOPs, since polymorphism is currently implicitly supported; the high level actions residing in the DL modules, as presented in Section 2.5, are polymorphic and can operate on a variety of types. Inheritance is also expected to allow designers to reuse digital object typing definitions. The concept of definition reuse through inheritance has been discussed in [8], although targeted on information retrieval enhancements.

Although DOPs are currently implemented atop the FEDORA repository, we believe that the presented concepts are of broader interest. The core type enforcement implementation of DOPs regarding digital object instances and their respective behavior is FEDORA independent and only stored digital object operations are tied to FEDORA specific functionality (e.g. `getDatastream`, `saveDatastream` services). Taken into consideration that DOPs, conceptually, relate to the OO paradigm and the digital object modeling approach of Kahn and Wilensky [9], we argue that there are strong indications that DOPs can be implemented in the context of other DL systems as well.

DOPs are complementary to FEDORA, or any other underlying repository. FEDORA can effectively handle low-level issues regarding digital object storage, indexing and retrieval. DOPs provide an architecture for the effective manipulation of digital objects in the higher level context of DL application logic. DOPs behaviors are divided into private and public, in order to support encapsulation, while their definition is performed in the object's respective prototype. FEDORA implements behaviors in terms of *disseminators*, which associate functionality with datastreams. FEDORA disseminators must be attached to each individual digital object upon ingestion time. With DOPs, all objects of the same type behave in the same manner; their respective behaviors are dynamically binded to the instances at runtime, while the behaviors are defined *once and in one place*, increasing management and maintenance capabilities. aDORe [4] deploys a behavior mechanism that, although it is similar to FEDORA, it attaches behaviors to stored digital objects in a more dynamic fashion, upon dissemination time, using disseminator-related rules stored in a knowledge base. Finally, DOPs behaviors operate on digital objects in a more fine-grained manner, since they can explicitly identify and operate upon the contents of FEDORA datastreams.

[5] enables the introspection of digital object structure and behavior. A DOP can be conceived as a meta-level entity that provides structural and behavioral metadata for a specific subset of base-level digital objects. Put in other terms,

a DOP acts as an introspection guide for its respective digital object instances. DOP supplied type conformance and type-driven introspection of digital object structure and behavior allows third parties to adjust to each object's "idiosyncrasy" in a uniform manner.

References

1. L. Cardelli and P. Wegner. On understanding types, data abstraction, and polymorphism. *ACM Computing Surveys*, 17(4):471–522, 1985.
2. W. Cook and J. Palsberg. A denotational semantics of inheritance and its correctness. In *Proceedings of the ACM Conference on Object-Oriented Programming: Systems, Languages and Application (OOPSLA)*, pages 433–444, New Orleans, Louisiana, USA, 1989.
3. *DCMI Metadata Terms.* Dublin Core Metadata Initiative, January 2005.
4. H. Van de Sompel, J. Bekaert, X. Liu, L. Balakireva, and T. Schwander. adore: A modular, standards-based digital object repository. *The Computer Journal*, 48(5):514–535, 2005.
5. N. Dushay. Localizing experience of digital content via structural metadata. In *Proceedings of the Joint Conference on Digital Libraries*, pages 244–252, Portland, Oregon, USA, 2002.
6. *Encoded Archival Description (EAD).* Library of Congress, 2006.
7. *Introduction to Fedora Object XML.* Fedora Project.
8. N. Fuhr. Object-oriented and database concepts for the design of networked information retrieval systems. In *Proceedings of the 5th international conference on Information and knowledge management*, pages 164–172, Rockville, Maryland, USA, 1996.
9. R. Kahn and R. Wilensky. *A Framework for Distributed Digital Object Services.* Corporation of National Research Initiative - Reston, VA, 1995.
10. *METS: An Overview & Tutorial.* Library of Congress, Washington, D.C., 2006.
11. *Resource Description Framework (RDF).* World Wide Web Consortium.
12. U.S Reddy. Objects as closures: Abstract semantics of object-oriented languages. In *Proceedings of the ACM Conference on Lisp and Functional Programming*, pages 289–297, Snowbird, Utah, USA, 1988.
13. K. Saidis, G. Pyrounakis, and M. Nikolaidou. On the effective manipulation of digital objects: A prototype-based instantiation approach. In *Proceedings of the 9th European Conference on Digital Libraries*, pages 26–37, Vienna, Austria, 2005.
14. T. Staples, R. Wayland, and S. Payette. The fedora project: An open-source digital object repository management system. *D-Lib Magazine*, 9(4), April 2003.
15. T. Staples, R. Wayland, and S. Payette. Using mpeg-21 dip and niso openurl for the dynamic dissemination of complex digital objects in the los alamos national laboratory digital library. *D-Lib Magazine*, 10(2), February 2004.

Design, Implementation, and Evaluation of a Wizard Tool for Setting Up Component-Based Digital Libraries

Rodrygo L.T. Santos, Pablo A. Roberto,
Marcos André Gonçalves, and Alberto H.F. Laender

Department of Computer Science, Federal University of Minas Gerais
31270-901 Belo Horizonte MG, Brazil
{rodrygo, pabloa, mgoncalv, laender}@dcc.ufmg.br

Abstract. Although component-based architectures favor the building and extension of digital libraries, the configuration of such systems is not a trivial task. Our approach to simplify the tasks of constructing and customizing component-based digital libraries is based on an assistant tool: a setup wizard that segments those tasks into well-defined steps and drives the user along these steps. For generality purposes, the architecture of the wizard is based on the 5S framework and different wizard versions can be specialized according to the pool of components being configured. This paper describes the design and implementation of this wizard, as well as usability experiments designed to evaluate it.

1 Introduction

The complexity of a digital library, with respect to its content and the range of services it may provide, varies considerably. As an example of a simple system, we could cite BDBComp (*Biblioteca Digital Brasileira de Computação*) [7], which provides, basically, searching, browsing, and submission facilities. More complex systems, such as CITIDEL (Computing and Information Technology Interactive Digital Educational Library) [3], may also include additional services such as advanced searching and browsing through unified collections, binding, discussion lists, etc.

Many of the existing digital libraries are based on monolithic architectures and their development projects are characterized by intensive cycles of design, implementation and tests [13]. Several have been built from scratch, aiming to meet the requirements of a particular community or organization [4].

The utilization of modular architectures, based on software components, beyond being a widely accepted software engineering practice, favors the interoperability of such systems at the levels of information exchange and service collaboration [13].

However, although component-based architectures favor the building and extension of digital libraries, the configuration of such systems is not a trivial task. In this case, the complexity falls on the configuration at the level of each component and on the resolution of functional dependencies between components.

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 135–146, 2006.

In existing systems, in general, such configurations are performed manually or via command-line scripts. Both alternatives, however, seem inappropriate in a broader context of digital libraries utilization. Instead, higher level techniques to support the creation of complete digital libraries in a simple manner should be investigated [14].

The approach taken in this paper for simplifying the tasks of constructing and customizing digital libraries consists in segmenting such tasks into steps and in driving the user along these steps. This approach is achieved through the development of a digital library setup wizard running on top of a pool of software components.

Wizards are applications specially suited for assisting users on the execution of both complex and infrequent tasks, presenting such tasks as a series of well-defined steps. Though efficient as assistant tools, such applications are not useful for didactical purposes; on the contrary, they should be designed to hide most of the complexity involved in the task to be accomplished. Besides, they should provide a supplementary rather than substitutive way to accomplish the task, so that they do not restrict its execution by specialist users [8].

This paper is organized as follows. In Section 2, the architecture of the wizard is described in details. Following, Section 3 shows some usage examples. In Section 4, we discuss the usability experimental evaluation of the prototype developed. Section 5 discusses related work. Finally, Section 6 presents conclusions and perspectives for future work.

2 Architecture Overview

In this section, we describe the architecture of the wizard, which basically follows the MVC (Model-View-Controller) framework [2] with the addition of a persistence layer.

The *model* layer was primarily designed [12] based on configuration requirements gathered from the ODL (Open Digital Libraries) framework [14]. Later, it was extended in order to support the configuration of different component pools. Such extension was conceived inspired on the definition of a digital library from the 5S (Streams, Structures, Spaces, Scenarios, Societies) framework [6]. Accordingly to 5S, a typical digital library is informally defined as a set of mathematical components (e.g., collections, services), each component being precisely defined as functional compositions or set-based combinations of formal constructs from the framework. Our configuration model was devised regarding the components that make up a 5S-like digital library as configurable instances of software components provided by a component pool. By "configurable instances", we mean software components whose behaviors are defined as sets of user-configurable parameters.

The class diagram [11] in Fig. 1 shows a simplified view of the devised model. As shown in the diagram, a *digital library* is implemented as a set of configurable instances of *provider* components, among those supplied by the *pool* being used. A provider may be typed either a *repository* or a *service*, according to its role within the library. For orthogonality purposes, the digital library itself is also implemented as a configurable instance of a *component*. Additionally, components may be declared mandatory, as well as dependent on other components. The configuration of each component is implemented as a set of *parameters*, semantically organized into parameter groups. For validation purposes, each parameter is associated to an existing Java type; they may also have a default value, in conformance with their defined type. Parameters may be also declared mandatory (not null) and/or repeatable (with cardinality greater than one).

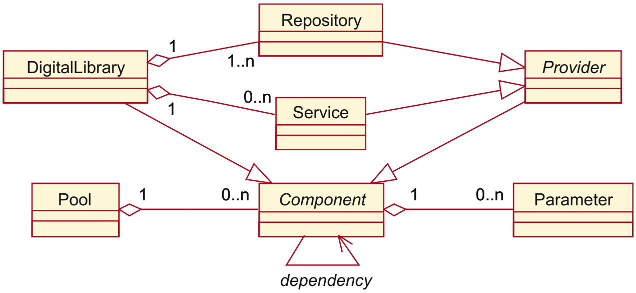


Fig. 1. Class diagram for the model layer

View and *controller* layers are integrated – a common simplification of the original MVC framework. They are responsible for handling user interactions, performing the corresponding modifications to the configuration model and displaying the updated model back to the user. Once user interactions are directly transmitted to the model, users modify a clone rather than the original configuration of each component. This allows users to cancel all the modifications performed to a given component at any time.

The configuration interface is organized into steps, in a wizard-like fashion. Each step comprises a major aspect of a digital library: the library itself, its collections and/or metadata catalogs, and the services it may provide. In each of these steps, the parameters associated to each of the components they list are presented in dynamically created, tab-organized forms (Figs. 2, 3, and 4). Each tab corresponds to a parameter group. Form elements are designed according to the type of the parameter they represent: repeatable parameters are shown as lists, parameters representing file descriptors present a file chooser dialog, parameters with values restricted to an enumerable domain are displayed as a combo box, strings and integers are simply shown as text fields. The semantics of every parameter is displayed as a tooltip near the parameter label. Type-checking is performed against every value entered by the user; in case of an erroneous value, a corresponding exception is raised and the user is notified about the error.

The *persistence* layer is responsible for loading and saving the components configuration. Besides that, it is up to this layer the tasks of setting environment variables and preparing databases that support the execution of some components. Its working scheme is based on two XML documents: a *pool descriptor* and a *configuration log*. The pool descriptor document details every component in the pool, including all configuration parameters associated to them. The description of each configuration parameter contains path entries of the form *document*:*xpath_expression* that uniquely locate the parameter in each of its source documents. Since some path entries are dependent on auto-detected or user-entered information, both only known at runtime (e.g., the base directory of the wizard and the current digital library identifier), the pool descriptor document also comprises a list of definitions to be used in path entries declaration. For example, in the listing below, the path entry for the "libraryName" parameter is declared relatively to the definitions "wizardHome" (auto-detected) and "libraryId" (user-entered). The other document, a configuration log, acts as a cache for the persistence layer. It comprises information about the currently configured digital libraries running in the server.

```
<component id="library" type="model.pool.library.DigitalLibrary">
  <group>
    <label>General Configuration</label>
    <parameter id="libraryName" type="java.lang.String" mandatory="yes">
      <path>
        #wizardHome/res/libs.xml:/config/library[@id='#libraryId']/name
      </path>
      <default>My New Library</default>
      <label>Library Name:</label>
      <description>A human readable name for the library.</description>
    </parameter>
    ...
```

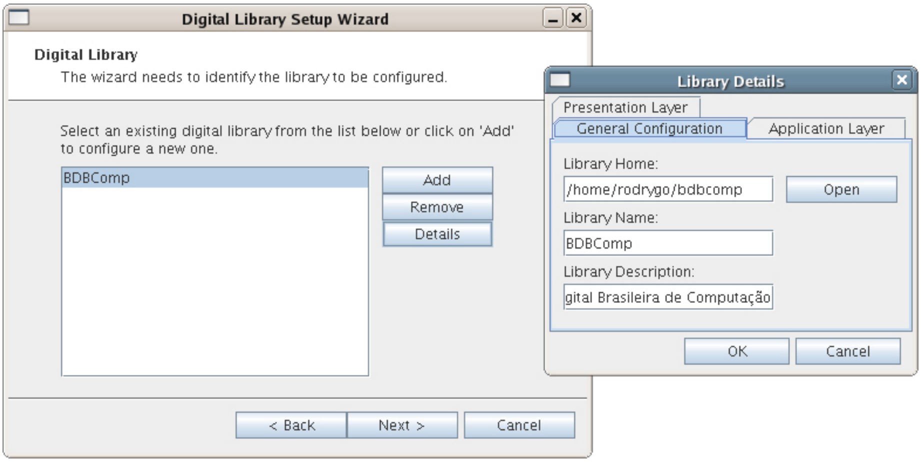
Both XML documents are handled via DOM. Loading and saving of components are performed through XPath expressions. Based on the specification of each component (from the pool descriptor document), configured instances of them are loaded into the digital library model; besides, a template of each component is added to the model so that new instances of components can be added later. Loading is performed in a lazy fashion, i.e., objects are created only when needed. On the other hand, saving is only performed at the end of the whole configuration task, as well as some additional tasks, such as environment variables and database setup, performed via system calls.

Specializing the wizard to assist the configuration of different component pools can be done just by providing a description document for each pool to be configured, as well as eventual accessory scripts for performing system calls. In fact, during the development project, we produced wizard versions for two component pools, namely, the ODL and WS-ODL frameworks.

3 Usage Examples

In this section, we show some usage examples of configuration tasks performed with the aid of the wizard developed.

The initial step welcomes the user and states the purpose of the wizard. The following step (Fig. 2) handles the digital library's configuration. At this step, previously configured digital libraries are listed by the wizard and the user can choose to modify or even remove any of them. Besides, he/she can choose to create a new digital library. Both library creation and modification are handled by a component editor dialog. For instance, selecting "BDBComp" from the list and clicking on "Details" opens this library's configuration editor dialog. This dialog comprises the digital library' general configuration (e.g., the library's home directory, name, and description), as well as its hosting information (e.g., the server name and port number for the library's application and presentation layers). Selecting a digital library from the list enables the "Next" button on the navigation bar.



Digital Library Setup Wizard

Digital Library

The wizard needs to identify the library to be configured.

Select an existing digital library from the list below or click on 'Add' to configure a new one.

BDBComp

Library Details

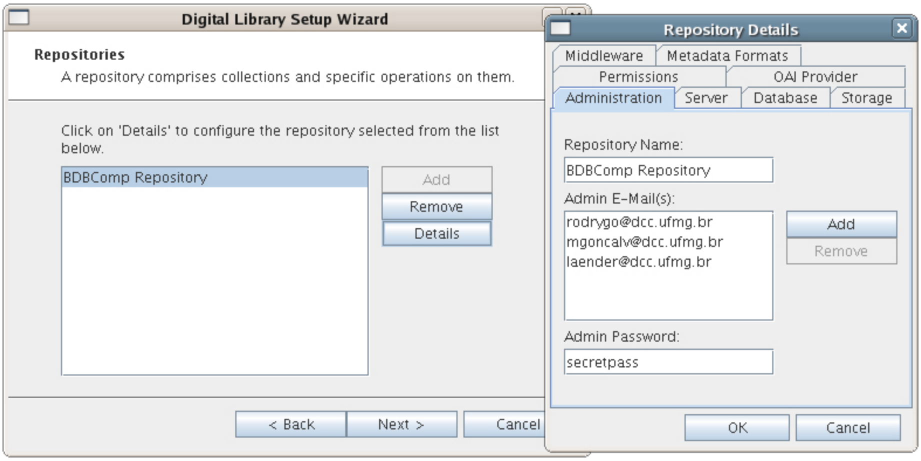
Library Home: /home/rodrygo/bdbcomp

Library Name: BDBComp

Library Description: gital Brasileira de Computação

Fig. 2. Configuring digital libraries

Clicking on "Next" drives the user to the following step (Fig. 3), which handles the configuration of the digital library's repositories. Similarly to the previous step, this one shows a list of existing repositories under the currently selected digital library so that the user can choose to modify or remove any of them. As in the previous step, he/she can also add a new repository to the library. Clicking on "Details" after selecting "BDBComp Repository" shows its configuration editor dialog. Repositories' configuration parameters include administrative data (e.g., repository administrators' e-mails and password), hosting information and access permissions (e.g., the repository's server name and a list of hosts allowed to access the repository), database connection and storage paths (e.g., the JDBC driver used to connect to the repository's database and the PID namespace associated to records stored in the repository), etc. Since the whole configuration is performed on the currently selected digital library and is only saved at the end of the configuration task, clicking on "Back" warns the user that selecting a new library to be configured implies discarding the current configuration. If there is at least one repository under the currently selected digital library, the "Next" button is enabled and the user can go forth.



Digital Library Setup Wizard

Repositories

A repository comprises collections and specific operations on them.

Click on 'Details' to configure the repository selected from the list below.

BDBComp Repository

Repository Details

Repository Name: BDBComp Repository

Admin E-Mail(s):
rodrygo@dcc.ufmg.br
mgoncalv@dcc.ufmg.br
laender@dcc.ufmg.br

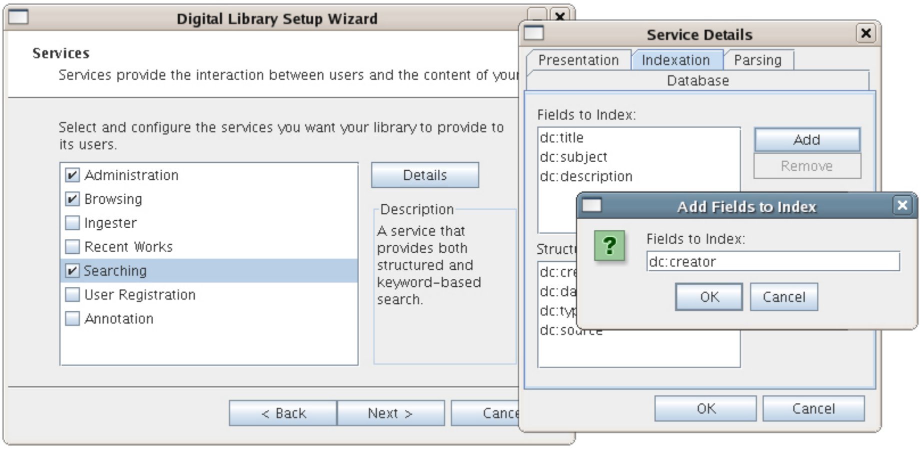
Admin Password: secretpass

Fig. 3. Configuring repositories

The following step (Fig. 4) handles the configuration of the digital library's services. A list of all the services provided by the pool of components being used is displayed – those already configured under the current library are marked. Selecting any of the services displays its description on the right panel. Trying to unmark a service which is an instance of a mandatory component raises an exception, as well as trying to mark a service component which depends on other components or to unmark a service component that other components depend on. Selecting a service component which has additional parameters to be configured enables the "Details" button. For instance, selecting "Browsing" and clicking on "Details" launches this service's configuration editor. Its configuration includes navigational parameters, such as a list of dimensions for browsing and the number of records to be displayed per page, and presentational parameters, such as the XSL stylesheets to be used when displaying browsing results. As another example, the "Searching" service's configuration includes parsing and indexation parameters, such as lists of delimiters, stopwords and fields to be indexed, among others.

After configuring the services that will be offered by the digital library, the user is driven to the penultimate step. This step summarizes all the configuration performed so far, showing a list of repositories and services comprised by the library being configured. If anything is wrong, the user can go back and correct the proper parameters. Otherwise, clicking on "Configure" saves the current digital library's configuration and drives the user to the last step.

The last step (Fig. 5) notifies the user about the result of the whole configuration task. If no problem has occurred while saving the configurations performed, links to the digital library's services are made available to the user.



Services

Services provide the interaction between users and the content of your

Select and configure the services you want your library to provide to its users.

- ☑ Administration
- ☑ Browsing
- ☐ Ingester
- ☐ Recent Works
- ☑ Searching
- ☐ User Registration
- ☐ Annotation

Description

A service that provides both structured and keyword-based search.

Fields to Index:

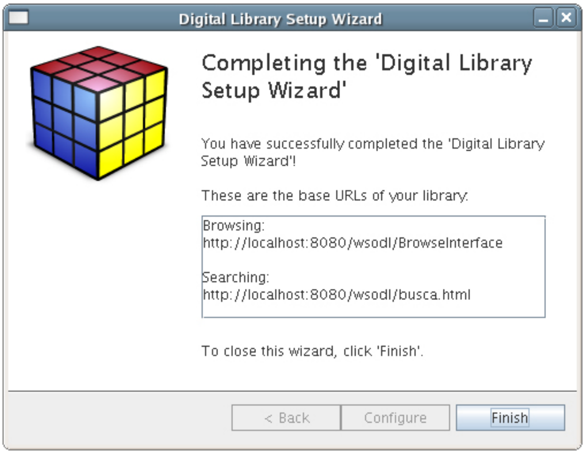
dc:title
dc:subject
dc:description

Struct

dc:cre
dc:da
dc:typ
dc:source

Fields to Index: dc:creator

Fig. 4. Configuring services



Completing the 'Digital Library Setup Wizard'

You have successfully completed the 'Digital Library Setup Wizard'!

These are the base URLs of your library:

Browsing:
http://localhost:8080/wsodl/BrowseInterface

Searching:
http://localhost:8080/wsodl/busca.html

To close this wizard, click 'Finish'.

Fig. 5. Configuration completion

4 System Evaluation

In order to evaluate the usability of our tool, we have conducted a series of experiments involving four users from Computer Science (CS) and four from Library and Information Science (LIS). The experiments included performing two configuration tasks and filling in an evaluation questionnaire. Both tasks highly explore all interface elements of the wizard, such as lists and file choosers. The first and simpler task, aimed at helping users to get familiar with the tool, consisted of modifying a few parameters of a pre-configured digital library. The second and more complex one consisted of configuring a whole library from scratch. Since the wizard prototype we tested was running on top of the WS-

ODL framework [10], we designed this second task to be comparable to the one performed at a command-line installation test conducted with that framework. Though data insertion is considered out of the scope of our tool but is performed in the command-line installation experiments of WS-ODL, the comparison was still possible since they measured the installation time at distinct checkpoints, allowing us to discard data insertion time while comparing the overall times. Table 1 shows the completion time and correctness from the two experiments conducted with the wizard prototype (namely, tasks #1 and #2), as well as those for the users who also performed the command-line driven configuration experiment (task #2^c). For comparison purposes, the performance of an expert user – the developers of the wizard and the WS-ODL framework – is also shown at the end of the table. Time is displayed in the form *hh:mm:ss* and correctness stands for the number of correctly executed items in the configuration task divided by the total number of items in that task.

Table 1. Completion time and correctness per task

	Completion Time			Correctness		
User	Task #1	Task #2	Task #2^c	Task #1	Task #2	Task #2^c
CS #1	00:05:16	00:10:48	–	1.00	1.00	–
CS #2	00:07:27	00:17:36	–	1.00	0.96	–
CS #3	00:07:26	00:08:09	01:36:00	1.00	1.00	0.78
CS #4	00:07:54	00:09:10	01:12:00	0.92	1.00	0.88
CS Mean	00:07:01	00:11:26	01:24:00	0.98	0.99	0.83
CS Std. Dev.	00:01:11	00:04:15	00:16:58	0.04	0.02	0.07
LIS #1	00:15:59	00:20:38	–	1.00	0.96	–
LIS #2	00:08:01	00:17:22	01:36:00	1.00	1.00	0.55
LIS #3	00:08:59	00:16:11	–	1.00	1.00	–
LIS #4	00:11:21	00:20:03	01:35:00	1.00	0.82	0.69
LIS Mean	00:11:05	00:18:33	01:35:30	1.00	0.95	0.62
LIS Std. Dev.	00:03:33	00:02:08	00:00:42	0.00	0.09	0.10
Global Mean	00:09:03	00:15:00	01:29:45	0.99	0.97	0.72
Global Std. Dev.	00:03:17	00:04:55	00:11:51	0.03	0.06	0.14
Expert	00:01:53	00:04:33	00:37:00	1.00	1.00	1.00

Comparing the wizard-guided and the command-line driven approaches for task #2 shows that configuring WS-ODL components with the aid of the wizard is much faster (about 500%, on average) than manually (hypothesis accepted by statistical analysis: t test with $\alpha = 0.05$). Configuration correctness is also substantially increased (about 34%, on average) with the aid of the wizard (hypothesis accepted by statistical analysis: t test with $\alpha = 0.05$). This is mainly due to the type-checking and component dependency checker systems of the wizard. Fastness and correctness attest the effectiveness of the wizard against the command-line driven approach. Effectiveness was also subjectively rated by users who participated in both tasks and measured based on a 5-point bipolar scale, ranging from 1 (worst rating) to 5 (best rating). On average, the effectiveness of the wizard-guided approach, in terms of easing the configuration task, was rated 4.5.

The learnability of the tool was also derived from Table 1. For such, we devised two measures: configuration efficiency and expertise. Efficiency stands for the total number of items in the task divided by the overall task completion time. Expertise measures how close the user's completion time is to the expert's completion time. Table 2 shows the values for these two learnability measures. Efficiency is measured in terms of task items performed per minute.

Table 2. Efficiency and expertise per task

	Efficiency			Expertise		
User	Task #1	Task #2	Task #2^c	Task #1	Task #2	Task #2^c
CS #1	2.47	2.59	–	0.36	0.42	–
CS #2	1.74	1.59	–	0.25	0.26	–
CS #3	1.75	3.44	0.93	0.25	0.56	0.39
CS #4	1.65	3.05	1.24	0.24	0.50	0.51
CS Mean	1.90	2.67	1.08	0.28	0.43	0.45
CS Std. Dev.	0.38	0.80	0.22	0.06	0.13	0.09
LIS #1	0.81	1.36	–	0.12	0.22	–
LIS #2	1.62	1.61	0.93	0.23	0.26	0.39
LIS #3	1.45	1.73	–	0.21	0.28	–
LIS #4	1.15	1.40	0.94	0.17	0.23	0.39
LIS Mean	1.26	1.52	0.93	0.18	0.25	0.39
LIS Std. Dev.	0.36	0.18	0.01	0.05	0.03	0.00
Global Mean	1.58	2.10	1.01	0.23	0.34	0.42
Global Std. Dev.	0.48	0.81	0.15	0.07	0.13	0.06
Expert	6.90	6.15	2.41	1.00	1.00	1.00

From Table 2, we can see that, in most cases (CS #2 and LIS #2 are the only exceptions), configuration efficiency is increased (about 33%, on average) from task #1 to task #2. Here we regard all task items as equally difficult, what is quite reasonable once all of them consist of setting configuration parameters. Also, the few items that differ in difficulty (e.g., choosing a file in a dialog or adding an item to a list) are homogeneously distributed across the two tasks. Expertise – another learnability indicator – is also increased (about 49%, on average) from task #1 to task #2, what could show that the wizard is easy to learn. However, the hypotheses of efficiency and expertise growth from task #1 to task #2 were rejected by statistical analysis (t test with $\alpha = 0.05$), what suggests that perhaps task #1 was not enough for users to become familiar with the tool.

From the questionnaire filled in by the users who performed the wizard-guided configuration tasks, we devised other two metrics: didactical applicability and satisfaction, both measured based on 5-point bipolar scales, ranging from 1 (worst rating) to 5 (best rating). On average, in terms of understanding of the concepts being configured (i.e., concepts pertaining to the domain of the component pool on top of which the wizard is running), the didactical applicability of the wizard was subjectively rated 3.75. This was an unexpected yet not unwelcome high value, since the design of wizards is not intended for didactical purposes. Satisfaction was measured in terms of comfort and ease of use. On average, users subjectively rated them 4.25 and 4, respectively.

5 Related Work

There are several works found in the literature that deal with component-based frameworks for building digital libraries. As far as we know, however, there are few works related specifically to the task of configuring such systems. In this section, we present four works that fall into the latter category.

5SGraph [15], a tool based on the 5S framework, provides a visual interface for conceptual modeling of digital libraries from a predefined metamodel. In the modeling task, the user interacts with the tool by incrementally constructing a tree where each node, picked from the metamodel, represents a construct of the digital library being modeled. Differently from the other works presented here, this one has a didactical goal: to teach the 5S theory.

BLOX [5] is a tool that hides most of the complexity involved in the task of configuring distributed component-based digital libraries. However, as occurs in 5SGraph, users interact with this tool in a flexible manner: its interface comprises a set of windows, each one representing the configuration of an ODL component.

The Greenstone suite [1] incorporates a wizard that allows non-specialist users to create and organize digital collections from local or remote documents. Driving the user step by step, this tool gets information such as the name and the purpose of the collection, administrator's e-mail, existing collections to serve as a model, base directories or URL's, etc. This tool, on the other hand, does not deal with the configuration of service provider components.

Finally, the OAIB application (Open Archives in a Box) [9], based on the COCOA framework (Components for Constructing Open Archives), provides a wizard for configuring metadata catalogs stored in RDBMS's. Its interface consists of a series of tabs where each tab presents different configuration options. Similarly to the wizard provided by the Greenstone suite, this one does not deal with the configuration of service providers.

Table 3 summarizes the characteristics of all these tools, comparing them to the ones present in our wizard.

Table 3. Wizard vs. related tools

	Wizard	5SGraph	BLOX	Greenstone	OAIB
task	configuration	modeling	configuration	configuration	configuration
objects	components	5S constructs	components	collections	catalogs
interaction	guided	flexible	flexible	guided	guided
didactical	no	yes	no	no	no

6 Conclusions and Future Work

This paper has presented a wizard tool for setting up component-based digital libraries. The tool is aimed at assisting users in the nontrivial task of configuring software components in order to build a fully functional digital library. The architecture of the wizard comprises a generic model layer for the purpose of supporting the configuration of different component pools upon minimal specialization.

The paper has also presented a usability experimental evaluation of a prototype running on top of the WS-ODL framework. Despite the relatively small number of users, the results (statistically meaningful) show that our approach is quite effective in easing the task of configuring that framework by hiding most of the complexity involved in the configuration task.

As future work, we plan to extend the wizard tool in order to support the customization of user interfaces and workflows. Though its comfort and ease of use have been well-rated, we plan to further enhance some interface aspects of the wizard based on users' suggestions and observations we made during the experiment sessions, in order to improve the overall learnability of the tool. Also, we intend to perform additional experiments in order to compare the guided and flexible interaction approaches, as provided by the wizard and the BLOX tool (for instance), respectively. In the near future, we plan to incorporate the wizard to the WS-ODL framework. Additionally, prototype versions for other component pools could be produced in order to test and expand the generality of the model layer.

Acknowledgments

This work was partially supported by CNPq funded projects I3DL and 5S/VQ. Authors would like to thank Allan J. C. Silva for his valuable help on statistical analysis of our experimental evaluation.

References

1. Buchanan, G., Bainbridge, D., Don, K. J., Witten, I. H.: A new framework for building digital library collections. In: Proceedings of the 5th ACM-IEEE Joint Conference on Digital Libraries (2005) 25–31
2. Burbeck, S.: Applications Programming in Smalltalk-80: How to use Model-View-Controller (MVC), tech. report. Softsmarts Inc. (1987)
3. CITIDEL. http://www.citidel.org, March (2006)
4. Digital Libraries in a Box. http://dlbox.nudl.org, March (2006)
5. Eyambe, L., Suleman, H.: A Digital Library Component Assembly Environment. In: Proceedings of the 2004 Annual Research Conference of the SAICSIT on IT Research in Developing Countries (2004) 15–22
6. Gonçalves, M. A., Fox, E. A., Watson, L. T., Kipp, N.: Streams, Structures, Spaces, Scenarios, Societies (5S): A Formal Model for Digital Libraries. ACM Transactions on Information Systems **22** (2004) 270–312

7. Laender, A. H. F., Gonçalves, M. A., Roberto, P. A.: BDBComp: Building a Digital Library for the Brazilian Computer Science Community. In: Proceedings of the 4th ACM-IEEE Joint Conference on Digital Libraries (2004) 23–24
8. MSDN. http://msdn.microsoft.com/library/en-us/dnwue/html/ch13h.asp, March (2006)
9. Open Archives in a Box. http://dlt.ncsa.uiuc.edu/oaib, March (2006)
10. Roberto, P. A.: Um Arcabouço Baseado em Componentes, Serviços Web e Arquivos Abertos para Construção de Bibliotecas Digitais. Master's thesis, Federal University of Minas Gerais (2006)
11. Rumbaugh, J., Jacobson, I., Booch, G.: The Unified Modeling Language Reference Manual. Addison-Wesley Professional (2004)
12. Santos, R. L. T.: Um Assistente para Configuração de Bibliotecas Digitais Componentizadas. In: I Workshop in Digital Libraries, Proceedings of the XX Brazilian Symposium on Databases (2005) 11–20
13. Suleman, H., Fox, E. A.: A Framework for Building Open Digital Libraries. D-Lib Magazine **7** (2001)
14. Suleman, H., Feng, K., Mhlongo, S., Omar, M.: Flexing Digital Library Systems. In: Proceedings of the 8th International Conference on Asian Digital Libraries (2005) 33–37
15. Zhu, Q., Gonçalves, M. A., Shen, R., Cassell, L., Fox, E. A.: Visual Semantic Modeling of Digital Libraries. In: Proceedings of the 7th European Conference on Digital Libraries (2003) 325–337

Design of a Digital Library for Early 20th Century Medico-legal Documents

George R. Thoma, Song Mao, Dharitri Misra, and John Rees

U.S. National Library of Medicine, Bethesda, Maryland, 20894, USA
{gthoma, smao, dmisra, jrees}@mail.nih.gov

Abstract. The research value of important government documents to historians of medicine and law is enhanced by a digital library of such a collection being designed at the U.S. National Library of Medicine. This paper presents work toward the design of a system for preservation and access of this material, focusing mainly on the automated extraction of descriptive metadata needed for future access. Since manual entry of these metadata for thousands of documents is unaffordable, automation is required. Successful metadata extraction relies on accurate classification of key textlines in the document. Methods are described for the optimal scanning alternatives leading to high OCR conversion performance, and a combination of a Support Vector Machine (SVM) and Hidden Markov Model (HMM) for the classification of textlines and metadata extraction. Experimental results from our initial research toward an optimal textline classifier and metadata extractor are given.

1 Introduction

As the United States moved from an agrarian economy to an industrial one during the late 19th and early 20th centuries, the need for food and drug regulation became increasingly important to American public health. Prior to this transformation, most food and medication came primarily from natural sources or trusted people, but as the nation's population became more urbanized, food and drug production became more of a manufacturing process. The mostly unregulated practice of adding chemicals and compounds and physical processes to increase the shelf life of foods, as well as outright medical quackery, became issues of political and social concern leading to legislation.

A landmark legislation, the 1906 Federal Food and Drug Act [1], established mechanisms for the federal government to seize, adjudicate, and punish manufacturers of adulterated or misbranded food, drugs and cosmetics. These federal activities were carried out by the various sub-offices we now know as the U.S. Food and Drug Administration (FDA). The legal proceedings associated with each case resulting from these activities were documented as *Notices of Judgment* (NJs), published synopses created on a monthly basis.

The U.S. National Library of Medicine (NLM) has acquired a collection of FDA documents (70,000+ pages) containing more than 65,000 NJs dating between 1906 and 1964. (In this paper, we refer to this collection as FDA documents.) To preserve these NJs and make them accessible, our goal is to create a digital archive of both page images and metadata. By providing access to NJs through metadata, this digital

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 147 – 157, 2006.

library will offer insight into U.S. legal and governmental history, but also into the evolution of clinical trial science and the social impact of medicine on health. The history of some of our best-known consumer items of today, such as Coca Cola, can be traced in the NJs. The intellectual value of this data for historians of medicine is expected to be high, and a Web service should increase its use exponentially.

Apart from providing access, digitization of this collection is needed for strictly preservation purposes since many of the existing volumes of NJs are one of a kind and the earliest ones are printed on paper that is extremely brittle and prone to crumbling. Constant physical handling of the print would probably shorten its lifespan considerably.

The creation of a digital library for this material requires a system for ingesting the scanned FDA documents, extracting the metadata, storage of documents (in TIFF and PDF forms) and metadata, and a Web server allowing access. This paper gives an overall system description (Section 2), and focuses on techniques for automated metadata extraction, experiments and results (Section 3).

2 System Description

A critical step in preserving the FDA documents for future access is the recording of the metadata elements pertaining to each NJ, and making the metadata accessible to users. The manual input of metadata for 65,000 NJs would be prohibitively expensive and error-prone. On the other hand, since these NJs are self-documenting, with important metadata elements (such as case number, description, defendant, judgment date), encoded in the pages following certain structured layout patterns, it is possible to consider automated extraction of these elements for a cost-effective and reliable solution. In our work, this automated metadata extraction is performed by using a prototype preservation framework called *System for the Preservation of Electronic Resources* (SPER) [2], which incorporates in-house tools to extract metadata from text-based documents through layout analysis.

SPER is an evolving Java-based system to research digital preservation functions and capabilities, including automated metadata extraction, retrieval of available metadata from Web-accessed databases, document archiving, and ensuring long term use through bulk file format migration. The system infrastructure is implemented through DSpace [3] (augmented as necessary to suit our purpose), along with a MySQL 5.0 database system.

The part of SPER that extracts metadata, called SPER-AME, is used for the preservation of the FDA documents. The overall workflow of the FDA documents through the system, as well as a description of the SPER-AME architecture with focus on components used for metadata extraction from the documents, are given below.

2.1 Preservation Workflow

Figure 1 depicts the high level workflow and processing steps involved in the preservation of the FDA documents. There are three basic steps, as described below.

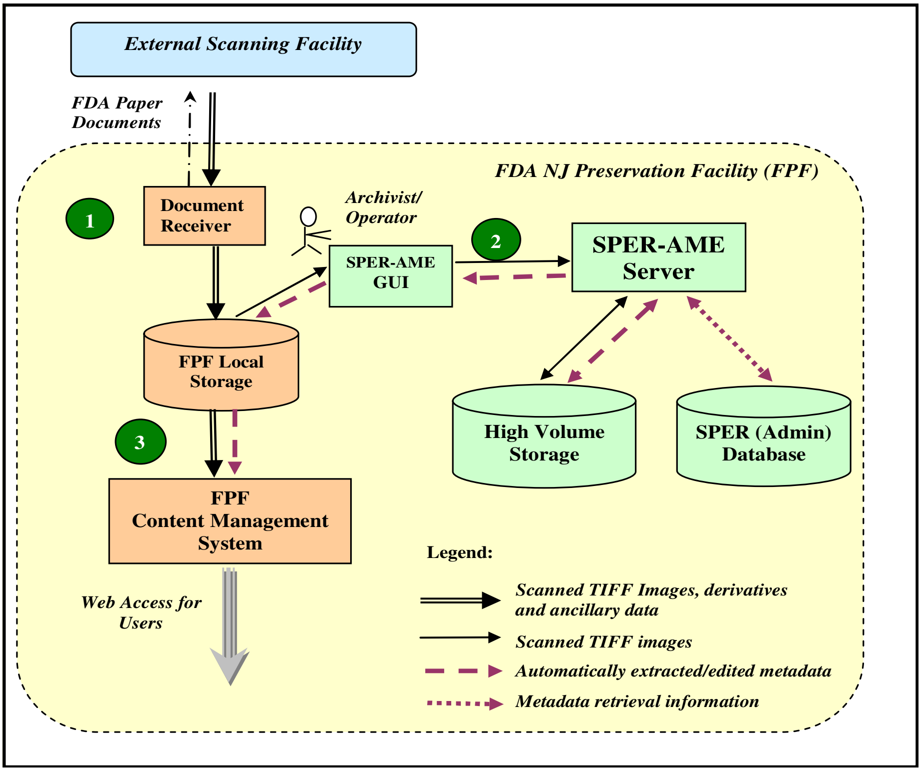

External Scanning Facility
FDA Paper Documents
FDA NJ Preservation Facility (FPF)
1
Document Receiver
Archivist/ Operator
SPER-AME GUI
2
SPER-AME Server
FPF Local Storage
3
High Volume Storage
SPER (Admin) Database
FPF Content Management System
Web Access for Users
Legend:
Scanned TIFF Images, derivatives and ancillary data
Scanned TIFF images
Automatically extracted/edited metadata
Metadata retrieval information

Fig. 1. Preservation Workflow for FDA Notices of Judgment

- As the first step, the FDA paper documents (either the originals, or, more frequently, their reproduction copies) are sent to a designated external scanning facility. The TIFF images of the scanned documents are sent back to an in-house facility (represented here as the *FDA NJ Preservation Facility* or FPF), and considered to be the master images for preservation. Besides these TIFF images, derivative documents such as PDF files, created for dissemination, are also received and stored at the FPF.
- In the next step, NJs are identified and metadata is automatically extracted from these TIFF documents using SPER-AME. In this client-server system, the back-end server process runs on a stand-alone Windows-2000 based server machine, while the frond-end client process, with a graphical user interface (GUI), runs on a console used by an archivist or operator.

  Using the SPER-AME GUI, the operator sends the master TIFF files in manageable batches to the server for automated extraction of metadata. The server receives the TIFF documents, identifies and extracts the embedded metadata for each NJ using the automated metadata extractor, stores both the image files and the extracted metadata (as XML files) in its storage system, and adds related information to the database. The operator may then view the extracted metadata for each NJ, perform editing if necessary, validate/qualify them for preservation, and download validated metadata to FPF local storage.

For efficiency, the SPER-AME server may perform metadata extraction from one batch while supporting interactive metadata review and editing by the operator from an already processed batch.

- In Step 3 the master TIFF images, the derivatives and the metadata are ingested to the FPF Content Management system for preservation and Web access. If necessary, the XML-formatted metadata from SPER will be reformatted to be compliant with the chosen Content Management system. This step will be discussed in a future report.

2.2 SPER-AME Architecture

As mentioned earlier, SPER is a configurable system, which (among other preservation functions) can accommodate metadata extraction for different types of documents and collections by using pluggable tailored interfaces encapsulating the internal characteristics of those documents. Here we describe a light-weight version of SPER (called SPER-AME), for the extraction of metadata from the FDA documents.

The SPER-AME system architecture is shown in Figure 2. Its operator interface runs as a separate GUI process, and communicates with the SPER-AME Server using Java Remote Method Invocation (RMI) protocols [4]. The File Copy Server is an RMI server, which runs on the operator's machine to transfer specified TIFF files from FPF local storage to the server upon request. These image files are stored on a multi-terabyte NetAPP RAID system and used for metadata extraction by the server. The three major components that participate in the metadata extraction process are the Metadata Manager, the Metadata Extractor, and the OCRConsole module. They are briefly described below. (Other essential components such as the Batch Manager and the Property Manager are not shown here for simplicity.)

Metadata Manager – This module receives all metadata-related requests from the GUI, through higher level RMI modules, and invokes lower level modules to perform the desired function such as extracting metadata from the documents, storing original/edited metadata in the database as XML files, and fetching these files to be sent to the operator upon request.

Metadata Extractor – This is the heart of the SPER-AME system, which identifies a specific NJ in a document batch and extracts the corresponding metadata elements by analyzing its layout from the associated OCR file. Further details on this module are provided in Section 3.

The metadata extractor for the FDA documents is chosen by the Metadata Manager (from a set of several extractors that have been developed for different document types) through an associated Metadata Agent module, shown in Figure 2. The Metadata Agent returns the metadata results from the Metadata Extractor in a standardized XML format.

OCRConsole– This is an *optical character recognition* module, external to SPER, invoked by the Metadata Extractor to take a TIFF image, generate a set of feature values for each character, such as its ASCII code, bounding box coordinates, font size, font attributes, etc., in the TIFF image, and store it in a machine-readable OCR output file. This OCR data is then used for layout analysis, metadata field classification, and metadata extraction.

The module *Metadata Validator*, shown in Figure 2, performs front-end checks such as missing mandatory metadata elements for an NJ item, invalid NJ identifiers, etc. so as to alert the FPF operator to review the item and make manual corrections as necessary.

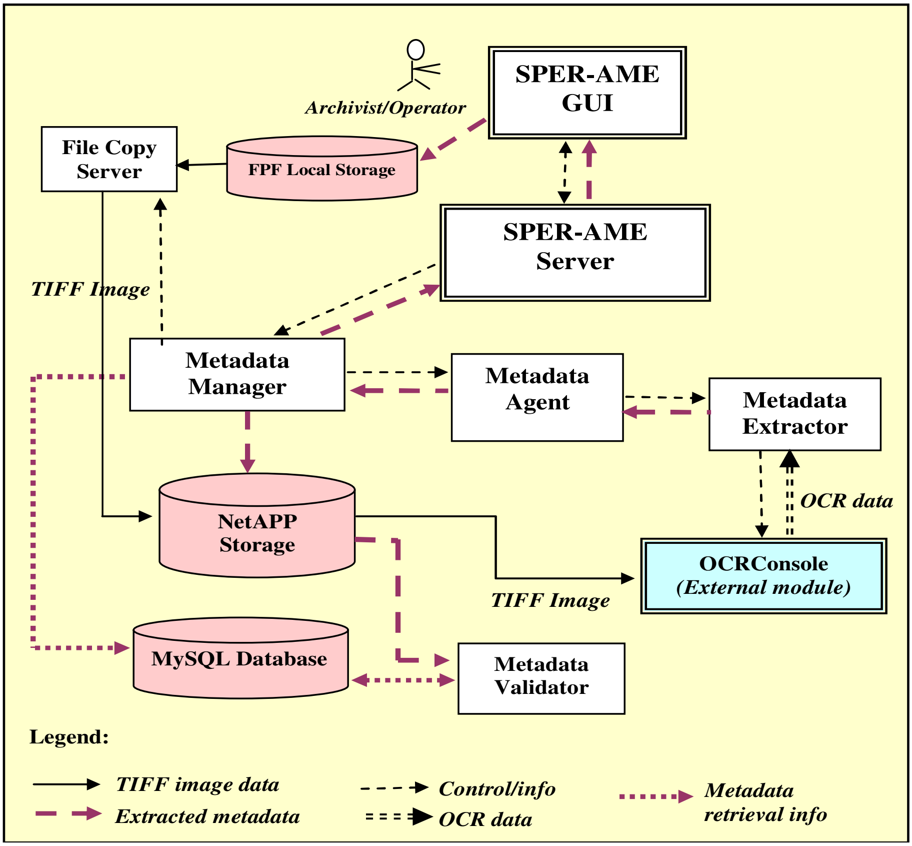


Fig. 2. SPER-AME System Components and Data Flow

3 Automated Metadata Extraction

Automated metadata extraction, an essential step in the economical preservation of these historic medico-legal documents, consists of the stages shown in Figure 3. Since the originals are brittle and have small font size, they are first photocopied at a magnified scale and appropriate toner level. Another reason for photocopying is the reluctance of sending one-of-a-kind rare documents to an outside facility. The photocopied version is then digitized as a TIFF image, which is recognized by the OCR-Console module whose design relies on libraries in a FineReader 6.0 OCR engine. Textlines are first segmented using the OCR output and then fourteen features are extracted from each textline. Layout is classified using layout type specific keywords. Each textline is classified as a case header, case body, page header (including page number, act name, and N. J. type or case range), and case category (e.g. cosmetics, food, drug, etc.) using a pre-trained layout type specific model file. Finally, metadata is extracted from the classified textline using metadata specific tags. Figure 4 shows an example of textline classes and its class syntax model that will be described in Section 3.2.

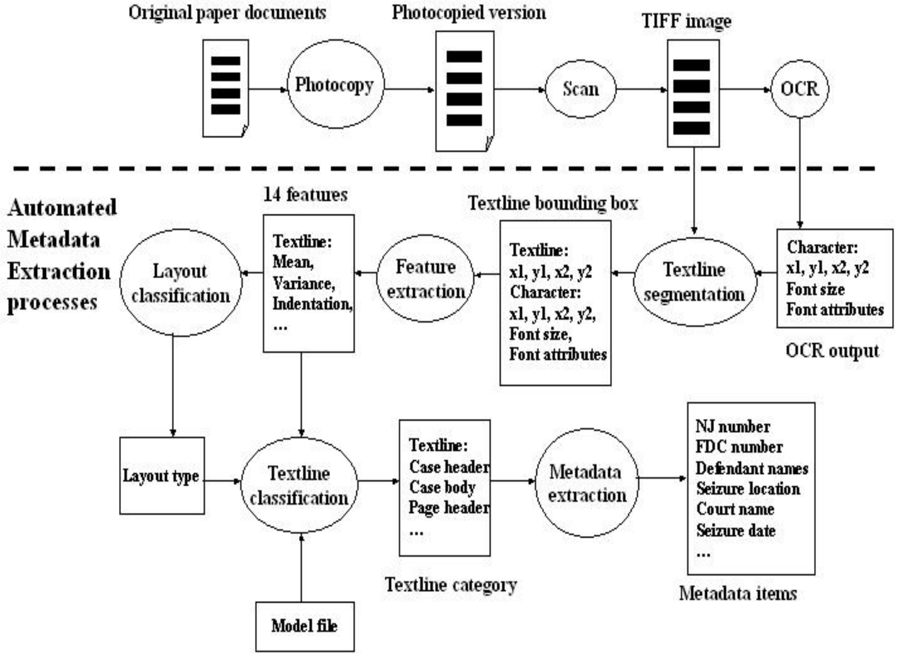
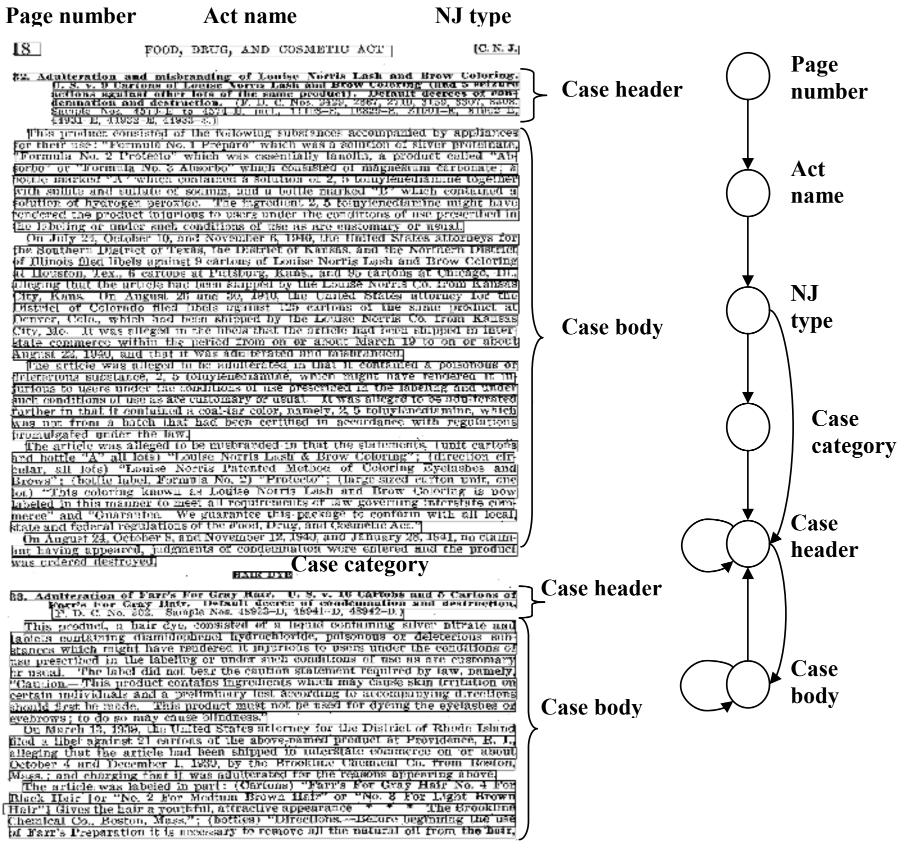

Original paper documents
Photocopy
Photocopied version
Scan
TIFF image
OCR
Automated Metadata Extraction processes
14 features
Textline bounding box
Layout classification
Textline: Mean, Variance, Indentation, ...
Feature extraction
Textline: x1, y1, x2, y2 Character: x1, y1, x2, y2, Font size, Font attributes
Textline segmentation
Character: x1, y1, x2, y2 Font size Font attributes
OCR output
Layout type
Textline classification
Textline: Case header Case body Page header ...
Metadata extraction
NJ number FDC number Defendant names Seizure location Court name Seizure date ...
Textline category
Metadata items
Model file

Fig. 3. Automated metadata extraction system. Ovals represent processes and rectangles represent objects or data.

In the following subsections, we first describe required metadata and layout classification, and then describe the 14 features extracted from each textline. Given next are the methods for classifying textlines, and metadata extraction from these classified textlines. Finally, we report experimental results.

3.1 Metadata and Layout Classification

Metadata important for future access to the FDA documents occur in the text. There are also metadata that are either constant such as format of the image (e.g., TIFF) or related to system operation (e.g., metadata creation time stamp). Table 1 provides a list of the metadata items of interest contained in these documents. Note that IS and Sample numbers are related to "Interstate Shipment" of food, drug and cosmetic



18 FOOD, DRUG, AND COSMETIC ACT [C. N. J.

32. Adulteration and misbranding of Louise Norris Lash and Brow Coloring. U. S. v. 9 Cartons of Louise Norris Lash and Brow Coloring (and 3 seizure actions against other lots of the same product). Default decrees of condemnation and destruction. (F. D. C. Nos. 2429, 2867, 2710, 3159, 3307, 3408. Sample Nos. 4511-E to 4571-E, incl., 11108-E, 10829-E, 21501-E, 31562-E, 41931-E, 41932-E, 41933-E.)

This product consisted of the following substances accompanied by appliances for their use: "Formula No. 1 Prepare" which was a solution of silver proteinate, "Formula No. 2 Protecto" which was essentially lanolin, a product called "Absorbo" or "Formula No. 3 Absorbo" which consisted of magnesium carbonate; a bottle marked "A" which contained a solution of 2, 5 toluylenediamine together with sulfite and sulfate of sodium, and a bottle marked "B" which contained a solution of hydrogen peroxide. The ingredient 2, 5 toluylenediamine might have rendered the product injurious to users under the conditions of use prescribed in the labeling or under such conditions of use as are customary or usual.

On July 24, October 10, and November 6, 1940, the United States attorneys for the Southern District of Texas, the District of Kansas, and the Northern District of Illinois filed libels against 9 cartons of Louise Norris Lash and Brow Coloring at Houston, Tex., 6 cartons at Pittsburg, Kans., and 95 cartons at Chicago, Ill., alleging that the article had been shipped by the Louise Norris Co. from Kansas City, Kans. On August 26 and 30, 1940, the United States attorney for the District of Colorado filed libels against 125 cartons of the same product at Denver, Colo., which had been shipped by the Louise Norris Co. from Kansas City, Mo. It was alleged in the libels that the article had been shipped in interstate commerce within the period from on or about March 19 to on or about August 22, 1940, and that it was adulterated and misbranded.

The article was alleged to be adulterated in that it contained a poisonous or deleterious substance, 2, 5 toluylenediamine, which might have rendered it injurious to users under the conditions of use prescribed in the labeling and under such conditions of use as are customary or usual. It was alleged to be adulterated further in that it contained a coal-tar color, namely, 2, 5 toluylenediamine, which was not from a batch that had been certified in accordance with regulations promulgated under the law.

The article was alleged to be misbranded in that the statements, (unit cartons and bottle "A" all lots) "Louise Norris Lash & Brow Coloring"; (direction circular, all lots) "Louise Norris Patented Method of Coloring Eyelashes and Brows"; (bottle label, Formula No. 2) "Protecto"; (large-sized carton unit, one lot) "This coloring known as Louise Norris Lash and Brow Coloring is now labeled in this manner to meet all requirements of law governing interstate commerce" and "Guarantee. We guarantee this package to conform with all local, state and federal regulations of the Food, Drug, and Cosmetic Act."

On August 24, October 8, and November 12, 1940, and January 28, 1941, no claimant having appeared, judgments of condemnation were entered and the product was ordered destroyed.

HAIR DYE

33. Adulteration of Farr's For Gray Hair. U. S. v. 10 Cartons and 5 Cartons of Farr's For Gray Hair. Default decree of condemnation and destruction. (F. D. C. No. 202. Sample Nos. 45923-D, 45941-D, 45942-D.)

This product, a hair dye, consisted of a liquid containing silver nitrate and tablets containing diamidophenol hydrochloride, poisonous or deleterious substances which might have rendered it injurious to users under the conditions of use prescribed in the labeling or under such conditions of use as are customary or usual. The label did not bear the caution statement required by law, namely, "Caution.—This product contains ingredients which may cause skin irritation on certain individuals and a preliminary test according to accompanying directions should first be made. This product must not be used for dyeing the eyelashes or eyebrows; to do so may cause blindness."

On March 13, 1939, the United States attorney for the District of Rhode Island filed a libel against 21 cartons of the above-named product at Providence, R. I., alleging that the article had been shipped in interstate commerce on or about October 4 and December 1, 1939, by the Brookline Chemical Co. from Boston, Mass.; and charging that it was adulterated for the reasons appearing above.

The article was labeled in part: (Cartons) "Farr's For Gray Hair No. 4 For Black Hair [or "No. 2 For Medium Brown Hair" or "No. 3 For Light Brown Hair"] Gives the hair a youthful, attractive appearance * * * The Brookline Chemical Co., Boston, Mass."; (bottles) "Directions.—Before beginning the use of Farr's Preparation it is necessary to remove all the natural oil from the hair,

Fig. 4. Textline classes in a sample TIFF image and its class syntax model

products and are used to identify a specific type of case. FDC and F&D numbers are used to categorize cases into Food, Drug and Cosmetic publications.

Table 1. Metadata items in historical medico-legal documents

Metadata item	Source
Case issue date	Page header text
Case/NJ number	Case header text
Case keyword	Case header text
F.D.C, Sample, IS and F&D numbers	Page header text or Case header text
Defendant Name(s)	Case body text
Adjudicating court jurisdiction	Case body text
Seizure location	Case body text
Seizure date	Case body text

These historical documents possess different layout types. Figure 5 shows three typical ones. We recognize the layout types by layout specific keywords from OCR results. For example, keywords such as "QUANTITY" and "LIBEL FILED" in layout type 1 are used for its detection. Once the layout type of a set of TIFF images is detected, a classification model is learned for this particular layout type, and used for textline classification in subsequent TIFF images possessing the same layout.

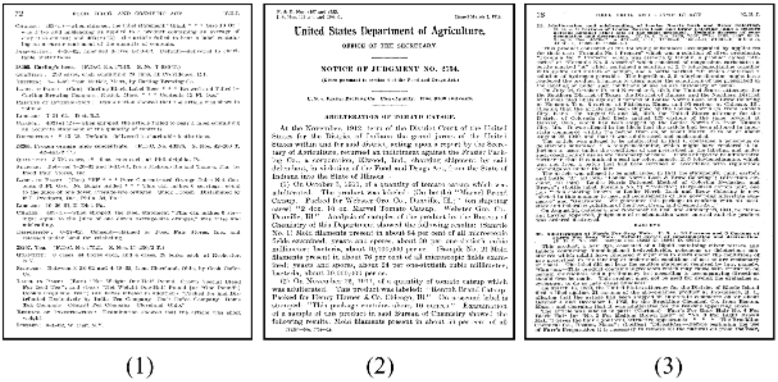


Fig. 5. Three typical layout types. Note that capitalized keywords such as "QUANTITY" and "NATURE OF CHARGE" are used to tag case body text in layout type 1, while case body text in layout types 2 and 3 appears as free text without such tags.

3.2 Features, Textline Classification and Metadata Extraction

We extract a set of 14 features from each textline using OCR results. They are 1: ratio of black pixels; 2-5: mean of character width, height, aspect ratio, and area; 6-9: variance of character width, height, aspect ratio, and area; 10: total number of letters and numerals/total number of characters; 11: total number of letters/total number of letters and numerals; 12: total number of capital letters/total number of letters; 13-14: indentation where 00 denotes center line, 10 denotes left indented line, 11 denotes full line, and 01 denotes right indented line, thus 13th feature value could indicate if the line touches the left margin, and 14th feature value could indicate if the line touches the right margin.

We classify textlines by a method that combines static classifiers with stochastic language models representing temporal class syntax. Support Vector Machines (SVMs) [5] are used as static feature classifiers. They achieve better classification performance by producing nonlinear class boundaries in the original feature space by constructing linear space in a larger and transformed version of the original feature space. However, they cannot model location evolution or class syntax as shown in Figure 4 in a sequence of class labels. On the other hand, stochastic language models such as Hidden Markov Models (HMMs) [6] are appropriate to model such class syntax. When features from different textline classes overlap in feature space, SVM classifiers could produce misclassification errors, while HMMs can correct such errors by enforcing the class syntax constraints. We therefore combine SVMs and HMMs in our algorithm [7] for optimal classification performance.

To represent class syntax in a one-dimensional sequence of labeled training textlines using HMM, we order textlines from left to right and top to bottom. Each distinct state in the HMM represents a textline class. State transitions represent possible class label ordering in the sequence as shown in Figure 4. Initial state probabilities and state transition probabilities are estimated directly from the class labels in the sequence. In the training phase, both the SVM and HMM are learned from the training dataset. In the test phase, they are combined in our algorithm [7] to classify textlines in the test dataset. Once a textline is classified, metadata items are extracted from it using metadata specific tags. Table 2 lists tag names used for different metadata items.

Table 2. Specific tags for metadata extraction

Metadata item	Tags
Case issue date	No tags needed (full text in identified field)
Case/NJ number	First word (in case header text)
Case keyword	Adulteration or misbranding (in case header text)
F.D.C, Sample, IS, and F&D numbers	Last open and closing parenthesis (in case header text)
Defendant Name(s)	Against, owned by, possession of, shipped by, manufactured by, transported by, consigned by
Adjudicating court jurisdiction	Filed in, convicted in, term of, session of, indictment in, pending in
Seizure location	From … to …
Seizure date	Shipped on or about, shipped during, shipped within the period

3.3 Experiments

To investigate optimal OCR and textline classification performance, we first photocopy the original document pages at different scales and toner levels, scan the photocopies into TIFF images, and then run our algorithm on these TIFF images. We select a scale of 130% for photocopying the 38 original pages of layout type 3 since this is the maximum possible scale that magnifies the text for the best OCR results while at the same time avoiding border cut-off. The classification algorithm is trained on a different training dataset of the same layout type at 130% scale and toner level 0. The reason for this choice is evident from Table 3 that shows the OCR performance (in terms of NJ number recognition error rate) and textline classification error rate at different toner levels. We consider an NJ number to be incorrectly recognized if any of its digits (up to five) is in error, or extra text is also included inadvertently. Test results are from an older version of the OCR engine. Upgrading this to the latest version is expected to significantly improve the character recognition accuracy.

Note that when toner level increases, there tends to be more noisy textlines and more misclassified textlines. When toner level decreases, text becomes too light and there are more OCR errors, and therefore fewer NJ numbers recognized correctly. OCR performance is optimal at toner level 0. Since misclassified textlines at toner level 0 is not very different from other toner levels, we select toner level 0 as the optimal value for our experiment. We can also see that the classification performance of our algorithm is relatively insensitive to the changes in toner level.

Table 3. Textline classification and OCR performance at different toner levels

Toner level (*Toner level increases from top to bottom)*	**Textline classification error rate** *(Number of incorrectly classified textlines/total number of textlines)*	**OCR performance (in terms of NJ number recognition error rate)** *(Number of incorrectly recognized NJ numbers/total number of NJ numbers)*
-3	2/2436	56/173
-2	0/2431	29/173
-1	2/2427	24/173
0	3/2436	22/173
+1	4/2437	26/173
+2	9/2476	26/173

We then train our classification algorithm on a training dataset of two of the layout types shown in Figure 5, and then test the algorithm on different test datasets of these layout types. We do not report experimental results for layout type 2 since it has very limited number of pages in our test sample. Table 4 shows the experimental results.

We see that textline classification errors from static classifiers (SVMs) are reduced by introducing class syntax models (HMMs) from 2.22% to 1.22% for layout type 1 and from 1.98% to 0.33% for layout type 3, a substantial improvement justifying our hybrid approach to the design of our classifier. Since most textlines are correctly classified, appropriate metadata items can be extracted from them using specific tags.

Table 4. Experimental results for two layout types

Layout type	**Training result**	**Test result**
1	Total pages: 30 Total textlines: 1,423 SVM errors: 5 SVM error rate: 5/1,423 = **0.35%** Corrected by HMM: 3 Final errors: 2 Final error rate: 2/1,423 = **0.14%**	Total pages: 189 Total textlines: 9,524 SVM errors: 211 SVM error rate: 211/9,524 = **2.22%** Corrected by HMM: 95 Final errors: 116 Final error rate: 116/9,524 = **1.22%**
3	Total pages: 30 Total textlines: 1,849 SVM errors: 3 SVM error rate: 3/1,849 = **0.16%** Corrected by HMM: 1 Final errors: 2 Final error rate: 2/1,849 = **0.11%**	Total pages: 195 Total textlines: 11,646 SVM errors: 231 SVM error rate: 231 / 11,646 = **1.98%** Corrected by HMM: 193 Final errors: 38 Final error rate: 38/11,646 = **0.33%**

4 Conclusion

In this paper, research toward a system for automated metadata extraction from historic medico-legal documents has been described. Specifically, a method that combines the power of static classifiers and class syntax models for optimal classification performance is introduced. In this method, each textline in these documents is classified into a category of interest. We tested our method on several hundred pages and show in our experimental results that the use of a class syntax model significantly reduces classification errors made by static classifiers. Future work includes automated selection of metadata specific tags for metadata extraction from free text, feature subset selection, and image enhancement during digitization.

Acknowledgment

This research was supported by the Intramural Research Program of the U.S. National Library of Medicine, National Institutes of Health.

References

1. Public Law 59-384, repealed in 1938 by 21 U.S.C. Sec 329 (a). And U.S Food and Drug Administration, “Federal Food and Drugs Act of 1906 (The "Wiley Act"),” http://www.fda.gov/opacom/laws/wileyact.htm (3 Feb. 2006).
2. Mao S, Misra D, Seamans J, Thoma, G. R.: Design Strategies for a Prototype Electronic Preservation System for Biomedical Documents, Proc. IS&T Archiving Conference, Washington DC, pages 48–53, (2005).
3. DSpace at MIT, http://www.dspace.org.
4. Java Remote Method Invocation, http://java.sun.com/products/jdk/rmi/.
5. Cortes C., Vapnik V.: Support-vector Network. Machine Learning. Vol. 20, pages 273-297, (1995)
6. Rabiner, L. R., Juang, B. H.: Fundamentals of Speech Recognition. Englewood Cliffs, NJ: Prentice-Hall. (1993).
7. Mao, S., Mansukhani, P., Thoma, G. R.: Feature Subset Selection and Classification using Class Syntax Models for Document Logical Entity Recognition. Proc. IEEE International Conference on Image Processing. Atlanta, GA, (2006). Submitted.

Expanding a Humanities Digital Library: Musical References in Cervantes' Works

Manas Singh[1], Richard Furuta[1], Eduardo Urbina[2], Neal Audenaert[1],
Jie Deng[1], and Carlos Monroy[1]

[1] Department of Computer Science, Texas A&M University
Center for the Study of Digital Libraries*
Texas A&M University
College Station, TX, 77843-3112
[2] Department of Hispanic Studies, Texas A&M University
Center for the Study of Digital Libraries*
Texas A&M University
College Station, TX, 77843-4238
cervantes@csdl.tamu.edu

Abstract. Digital libraries focused on developing humanities resources for both scholarly and popular audiences face the challenge of bringing together digital resources built by scholars from different disciplines and subsequently integrating and presenting them. This challenge becomes more acute as libraries grow, both in terms of size and organizational complexity, making the traditional humanities practice of intensive, manual annotation and markup infeasible. In this paper we describe an approach we have taken in adding a music collection to the *Cervantes Project*. We use metadata and the organization of the various documents in the collection to facilitate automatic integration of new documents—establishing connection from existing resources to new documents as well as from the new documents to existing material.

1 Introduction

As a digital library grows in terms of both size and organizational complexity, the challenge of understanding and navigating the library's collections increases dramatically. This is particularly acute in scenarios (e.g., scholarly research) in which readers need and expect to be able to survey all resources related to a topic of interest. While large collections with a rich variety of media and document sources make valuable information available to readers, it is imperative to pair these collections with tools and information organization strategies that enable and encourage readers to develop sophisticated reading strategies in order to fully realize their potential [11]. Traditional editorial approaches have focused on detailed hand editing—carefully reading and annotating every line on every page with the goal of producing a completed, authoritative edition. Often, such approaches are infeasible in a digital library environment. The sheer magnitude of many digital collections (e.g., the Gutenberg Project [13], the *Christian Classics Ethereal Library* [20], the *Making of America* [7][17]) make detailed hand editing unaffordably labor intensive, while the very nature of the project often conflicts with the traditional goal of producing a final,

* Authors' academic affiliations.

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 158 – 169, 2006.

fixed edition. Previously, we have described the multifaceted nature of humanities collections focused on a single author and argued that these projects will require automatic integration of many types of documents, drawn from many sources, compiled by many independent scholars, in support of many audiences [1]. Such collections are continuously evolving. As each new artifact is added to the collection, it needs to be linked to existing resources and the existing resources need to be updated to refer to the new artifact, where appropriate. Constructing these collections will require new tools and perspective on the practice of scholarly editing [10]. One such tool class is that supporting automatic discovery and interlinking of related resources.

The *Cervantes Project* [25] has been focused during the last ten years on developing on-line resources on the life and works of Miguel de Cervantes Saavedra (1547 – 1616), the author of *Don Quixote* [5], and thus has proven to be a rich environment for exploring these challenges. Given its canonical status within the corpus of Hispanic literature and its iconic position in Hispanic culture, the *Quixote* has received a tremendous amount of attention from a variety of humanities disciplines, each bringing its own unique questions and approaches. Within the broad scope of this project, individual researchers have made a variety of contributions, each centered on narrowly scoped research questions. Currently, work in the project can be grouped into six sub-projects: bibliographic information, textual studies, historical research, music, *ex-libris*, and textual iconography. Together, these contributions span the scope of Cervantes' life and works and their impact on society.

In this paper, we describe the approach that we have taken in connection with the presence and influence of music in Cervantes' works. The data for this project was collected by Dr. Juan José Pastor as part of his dissertation work investigating Cervantes' interaction with the musical culture of his time and the subsequent musical interpretations of his works [18]. Pastor's collection is organized in five main categories (instruments, songs, dances, composers, and bibliographical records) and contains excerpts from Cervantes' writings, historical and biographical information, technical descriptions, images, audio files, and playable scores from songs. Although Pastor has completed his dissertation, the collection is still growing, as new scores, images, and documents are located. For example, a recent addition, published in conjunction with the 400th anniversary of the publication of the *Quixote*, is a professionally-produced recording of 22 of the songs referred to by Cervantes [12].

The music sub-project reflects many aspects of the complexity of the *Cervantes Project* as a whole, and thus provides an excellent testbed for developing tools and strategies for integrating an evolving collection of diverse artifacts for multiple audiences. A key challenge has been determining how to build an interface that effectively integrates the various components, in a manner that supports the reader's understanding of the implicit and explicit relationships between items in the collection. In particular, since the collection is growing with Pastor's ongoing research, it was necessary that the interface be designed so that new resources could be easily added and the connections between new and old resources generated automatically. To address this challenge we have developed an automatic linking system that establishes relationships between resources based on the structural organization of the collection and various metadata fields associated with individual documents. An editor's interface allows users an easy way to add new resources to the collection and to specify the minimal set of metadata required to support link generation. Further, a reader's interface is provided that identifies references within texts to other items in the collection and dynamically generates navigational links.

2 Background

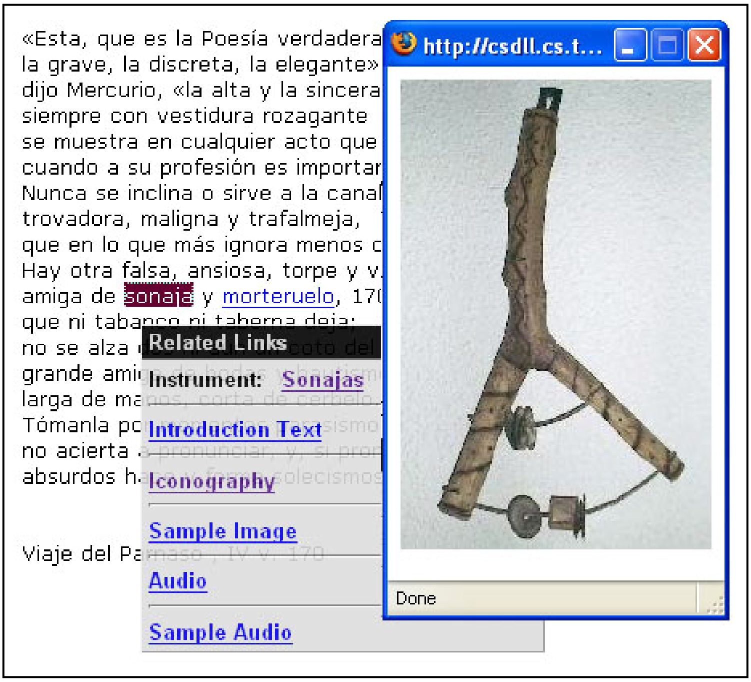
Developing a system to integrate resources within the collection required attention to three basic questions: What types of reader (and writer/editor) interactions are to be supported? What types of information and connections are to be identified? How will that information be identified and presented to readers? A brief survey of related projects will help to set the context for the design decisions we have made in these areas.

The *Perseus Project* [26] has developed a number of sophisticated strategies for automatically generating links in the context of cultural heritage collections [8][9]. Our work has been heavily influenced by their use of dense navigational linking both to support readers exploring subjects with which they are unfamiliar and to encourage readers more closely acquainted with a subject to more fully explore and develop their own interpretive perspectives. Early work focused on developing language based tools to assist readers of their extensive Greek and Latin collections. These tools linked words to grammatical analysis, dictionaries and other linguistic support tools, helping a wider audience understand and appreciate them. More recently, they have focused on applying some of the techniques and technologies developed for their Classical collection to a variety of other, more recent data sets including American Civil War and London collections. This work has focused on identifying names, places, and dates to provide automatically generated links to supplementary information and to develop geospatial representations of the collection's content. They have had good results from a layered approach using a combination of *a priori* knowledge of semi-structured documents (e.g., of the *British Directory of National Biography* and *London Past and Present*), pattern recognition, name entity retrieval, and gazetteers to identify and disambiguate references to people, places, and events.

A key technology for supporting this type of integration between resources within a collection is the use of name authority services. The SCALE Project (Services for a Customizable Authority Linking Environment) is developing automatic linking services that bind key words and phrases to supplementary information and infrastructure to support automatic linking for collections within the National Science Digital Library [19]. This collaborative effort between Tufts University and Johns Hopkins University builds on the tools and techniques developed in the Perseus Project in order to better utilize the authority controlled name lists, thesauri, glossaries, encyclopedias, subject hierarchies and object catalogs traditionally employed in library sciences in a digital environment.

As an alternative to authority lists, the Digital Library Service Integration (DLSI) project uses lexical analysis and document structure to identify anchors for key terms within a document [6]. Once the anchors are identified, links are automatically generated to available services based on the type of anchor and the specified rules. For example, if a term is a proper noun it can be linked to glossaries and thesauri to provide related information.

Also of relevance is the long history in the hypertext research community of link finding and of link structures that are more than simple source to destination connections. Early work in link finding includes Bernstein's Link Apprentice [4] and Salton's demonstration of applications [22] of his Smart system's vector-space model [21]. Link models related to our work include those that are multi-tailed, for example MHMT [15] and that represented in the Dexter model [14].



«Esta, que es la Poesía verdadera
la grave, la discreta, la elegante»
dijo Mercurio, «la alta y la sincera
siempre con vestidura rozagante
se muestra en cualquier acto que
cuando a su profesión es importar
Nunca se inclina o sirve a la canal
trovadora, maligna y trafalmeja,
que en lo que más ignora menos c
Hay otra falsa, ansiosa, torpe y v
amiga de sonaja y morteruelo, 17(
que ni tabanco ni taberna deja;
no se alza
grande ami
larga de ma
Tómanla pc
no acierta
absurdos h

Viaje del Pa

Fig. 1. Related links and a sample image for *sonaja* instrument

3 Interface and Usage Scenario

Within the context of the Cervantes music collection, we have chosen to focus on identifying interrelationships between the structured items in our collection in order to provide automatic support for the editorial process rather than relying on authority lists or linguistic features to connect elements of the collection to externally supplied information sources (such support for this could be added later, if warranted). We have divided the resources in our collection into categories of structured information (e.g., instruments, songs, composers). Each category contains a set of items (e.g., a particular song or composer). Each item is in turn represented by a structured set of documents. How the documents for any given item are structured is determined by the category it is a member of. For example, *arpa* (a harp) is an item within the instruments category. This instrument (like all other instruments) may have one or more of each of the following types of documents associated with it: introductory articles, images, audio recordings, historical descriptions, bibliographic references, links to online resources, and excerpts from the texts of Cervantes that refer to an *arpa*.

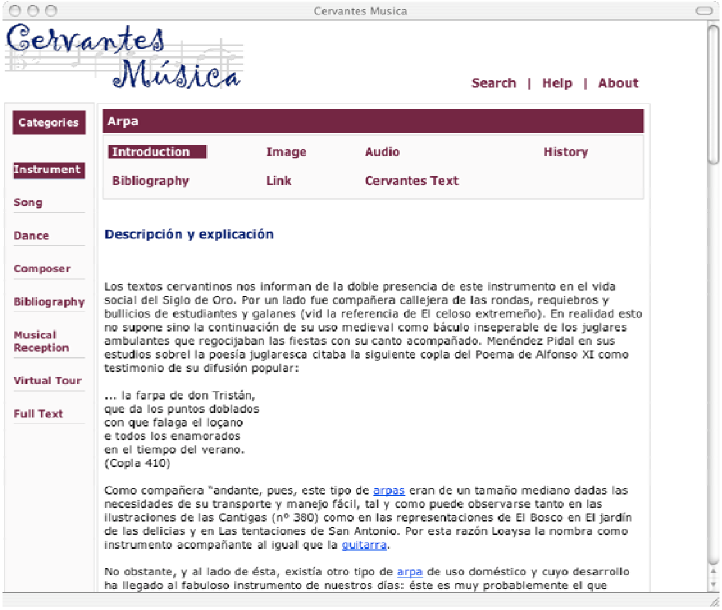
Each item is identified by its name and by a list of aliases. Our system identifies the references to these terms in all of the resources located elsewhere in the collection, either as direct references or within the metadata fields of non-textual documents. At present, the matching algorithm is a simple match between the longest-length term string found at the target. Once identified, the references are linked to the item.

The presentation of information to the reader uses a custom representation of links. This is because of the complexity of the object linked to—a complexity that reflects the multiple user communities that we expect will make use of the collection. Moreover, the collection provides multiple roots that reflect different reader specializations.

In developing the Cervantes music collection we have focused our design on meeting the needs of two primary communities of readers. One group is composed of Cervantes scholars and music historians interested in research about Cervantes' works and music. The second group is composed of non-specialists interested in gaining access to information they are unfamiliar with. For both the specialist and the non-specialist, the collection provides two major focal points, or roots, for access. For example, a reader might approach the music collection from the texts of Cervantes (which themselves compose a distinct collection), asking how a particular passage reflects Cervantes' understanding of contemporary musical trends or in order to better understand what, for example, an *albogue* looks and sounds like.[1] Another reader might begin by considering a particular composition that alludes to Cervantes and ask how this particular piece reflects (or is distinct from) other popular interpretations of the *Quixote*. Similarly, a non-expert might find his understanding of a particular opera enhanced by learning more about an obscure reference to one of Cervantes' works. In this way the linkages generated between these two distinct but related collections allow readers access to a rich and diverse body of resources from multiple perspectives to achieve a variety of goals. We refer to collections that exhibit this type of structure as being multi-rooted. Natural roots for the music collection include compositions (e.g., songs and dances), composers, instruments, and the writings of Cervantes. In the remainder of this section we present several brief reader interaction scenarios to help illustrate the design of the system from a reader's perspective. In the following section we present an overview of the technical design and implementation of the link generation system and the interface.

In the first scenario, a native, modern Spanish speaker is reading a less well-known text of Cervantes, *Viaje del Parnaso* (1614), and encounters a reference to an instrument she is unfamiliar with, the *sonaja*. Curious, she clicks on the link and a drop-down menu appears displaying links to the various types of documents present in the collection. She elects to view the 'sample image,' resulting in the display shown in Figure 1. The image sparks her curiosity and she decides to see what it sounds like by clicking on the 'sample audio' link. What is this, who would use it, and why? To find out more, she clicks to read the introductory text and finds a list of definitions where she learns that it is a small rustic instrument that was used in the villages by beating it against the palm of the hands. Interestingly, the Egyptians used it in the celebrations and sacrifices to the goddess. Having learned what she wanted to know, she returns to reading *Viaje del Parnaso*.

[1] "What are albogues?" asked Sancho, "for I never in my life heard tell of them or saw them." "Albogues," said Don Quixote, "are brass plates like candlesticks that struck against one another on the hollow side make a noise which, if not very pleasing or harmonious, is not disagreeable and accords very well with the rude notes of the bagpipe and tabor. [Chapter 65, Part 2, *Don Quixote]*.



Descripción y explicación

Los textos cervantinos nos informan de la doble presencia de este instrumento en el vida social del Siglo de Oro. Por un lado fue compañera callejera de las rondas, requiebros y bullicios de estudiantes y galanes (vid la referencia de El celoso extremeño). En realidad esto no supone sino la continuación de su uso medieval como báculo inseperable de los juglares ambulantes que regocijaban las fiestas con su canto acompañado. Menéndez Pidal en sus estudios sobrel la poesía juglaresca citaba la siguiente copla del Poema de Alfonso XI como testimonio de su difusión popular:

... la farpa de don Tristán,
que da los puntos doblados
con que falaga el loçano
e todos los enamorados
en el tiempo del verano.
(Copla 410)

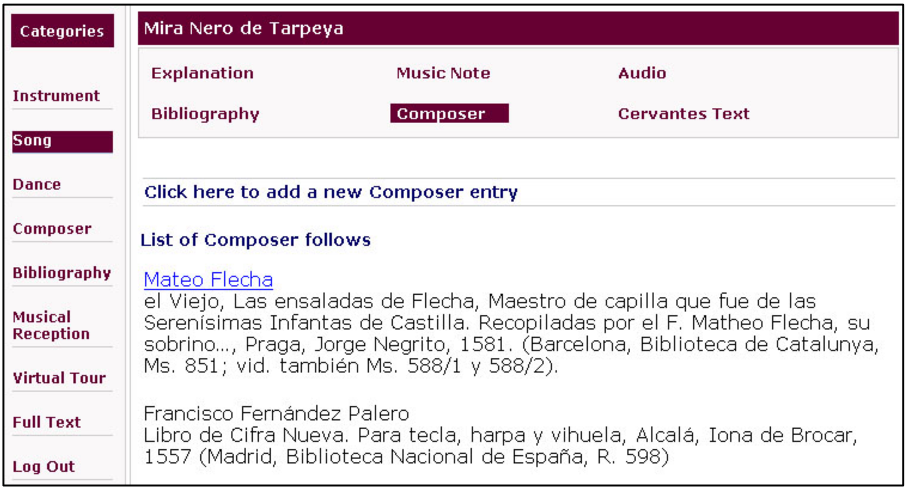
Como compañera "andante, pues, este tipo de arpas eran de un tamaño mediano dadas las necesidades de su transporte y manejo fácil, tal y como puede observarse tanto en las ilustraciones de las Cantigas (nº 380) como en las representaciones de El Bosco en El jardín de las delicias y en Las tentaciones de San Antonio. Por esta razón Loaysa la nombra como instrumento acompañante al igual que la guitarra.

No obstante, y al lado de ésta, existía otro tipo de arpa de uso doméstico y cuyo desarrollo ha llegado al fabuloso instrumento de nuestros días: éste es muy probablemente el que

Fig. 2. Learning more about the *arpa*

In the second scenario, a music historian with relatively little familiarity with *Don Quixote* or the other works of Cervantes is interested in exploring how string instruments were used and their societal reception. On a hunch, he decides to see how societal views of the harp and other instruments might be reflected in the works of Cervantes. Browsing the collection, he navigates to the section for the harp and peruses the texts of Cervantes that refer to the harp (Figure 2). After surveying that information, he explores some of the other instruments in order to get a broader perspective on how Cervantes tends to discuss and incorporate musical instruments in his writings. He finds a couple of passages that help to illustrate the ideas he has been developing, and makes a note of them to refer to later.

In the final scenario, an editor is working with the collection, adding the historical documents to the song, "Mira Nero de Tarpeya." As shown in Figure 3, he browses to the list of composers and notices that, while there is a link to Mateo Flecha, there is no information provided for Francisco Fernández Palero. He quickly navigates to the "composers" category, adds Palero as a new composer (Figure 4), and writes a short description of him and his relevance to classical music. The system recognizes the new composer and updates its generated links accordingly. Currently, since only minimal information is present, these links refer only to the newly written introductory text. A few weeks later, the editor returns to the collection after finding images, lists of songs written, and historical descriptions. He adds these through forms similar to the one he used to add Fernández Palero. Links to these new resources are now added to the drop down menu associated with references to Fernández Palero. In this way, the editor is able to focus on his area of expertise in finding and gathering new information that will enhance the scholarly content of the collection, removing the burden of manually creating links from all the existing documents to the newly added composer.



Mira Nero de Tarpeya

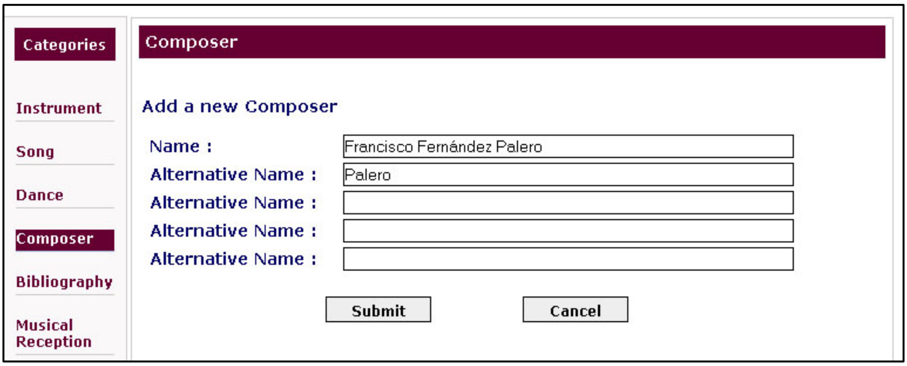
Click here to add a new Composer entry

List of Composer follows

Mateo Flecha
el Viejo, Las ensaladas de Flecha, Maestro de capilla que fue de las Serenísimas Infantas de Castilla. Recopiladas por el F. Matheo Flecha, su sobrino..., Praga, Jorge Negrito, 1581. (Barcelona, Biblioteca de Catalunya, Ms. 851; vid. también Ms. 588/1 y 588/2).

Francisco Fernández Palero
Libro de Cifra Nueva. Para tecla, harpa y vihuela, Alcalá, Iona de Brocar, 1557 (Madrid, Biblioteca Nacional de España, R. 598)

Fig. 3. Browsing a Song in the Editor's Interface



Composer

Add a new Composer

Name : Francisco Fernández Palero
Alternative Name : Palero
Alternative Name :
Alternative Name :
Alternative Name :

Fig. 4. Adding the composer Francisco Fernandez Palero

4 Organization of the Digital Library

Information in the collection is organized as hierarchical groups. At the highest level, materials are grouped into eight categories:

1 Instruments: information pertaining to the different musical instruments that have been referred to by Cervantes in his works.
2 Songs: information regarding the different songs that have influenced Cervantes.

3 Dances: resources related to the dances that have been referred to in Cervantes' texts.
4 Composers: the composers who have influenced Cervantes and his work.
5 Bibliography: bibliographical entries related to instruments, songs, and dances that have been referred to in Cervantes' texts.
6 Musical Reception: bibliographical entries about musical compositions that have been influenced by Cervantes or refer to his works.
7 Cervantes Text: full texts of Cervantes' works.
8 Virtual Tour: links to virtual paths, constructed and hosted using Walden's Paths [23]. This allows the information to be grouped and presented in different manners, catering to the interests of diverse scholars, thus opening up the digital library to unique interpretive perspectives.

Most categories are further subdivided into items. An item defines a unique logical entity, as appropriate for the type of category. For example, the category "Instruments" contains items such as *arpa* and *guitarra*. Similarly, each composer would be represented as an item in the respective category as would each dance and each song. The item is identified by its name, perhaps including aliases (e.g., variant forms of its name).

Artifacts associated with each item are further categorized into different topics like image, audio, and text. The topics under an item depend on the category to which the item belongs to. For example, an item under category "Instruments" will have topics like introduction, audio, image, text, and bibliography but an item under the category "Composer" will have topics like life, image, work, and bibliography.

An artifact (e.g., an individual picture; a single essay) is the "atomic" unit in the collection. Thus artifacts are grouped under topics, which in turn are grouped into items, which in turn are grouped into categories. A unique item identifier identifies each item in the digital library. Additionally, each artifact placed under an item is assigned a sub-item identifier that is unique among all the artifacts under that item. Thus all the artifacts, including texts, audio files, images, musical scores, etc., are uniquely identified by the combination of item identifier and sub-item identifier.

5 Interlinking

The process of creating interlinks and presenting the related links can be broadly classified into four major steps. The first is maintaining the list of item names for which information exists in the digital library. The second is a batch job, which identifies the reference of these terms in all the texts present in the digital library. The third step is a run time process, which, while displaying a text, embeds the terms that need to be linked with a hyperlink placeholder (i.e., hyperlink without any specific target). This step uses the data from the batch job to identify the terms that should be presented with the hyperlink for any text. The final step generates the actual related links for a term and is invoked only when the user clicks on a hyperlink placeholder. A description of these steps follows.

Maintaining the keyword list: In order for the system to provide related links, it should be able to identify the terms for which information exists in the digital library. This is achieved by maintaining a keyword list. To identify the variation in names a synonym list is also maintained. The system depends on the user to provide a list of synonyms for the item being added. This may include alternate names for the item or just variations in the spelling of the item name.

When a new item is added to the digital library its name or title is added to the keyword list and its aliases to the synonym list. In the following sections the keyword and synonym lists will be referred to collectively as keywords.

Document keyword mapping batch job: The document keyword mapping is created by indexing all the texts using Lucene and finding the references of each term in the keyword list among all the texts. This is done offline using a batch process. This also populates a document keyword map that maps each document to all the keywords it refers.

Runtime display of texts with hyperlink placeholders: While displaying a text the system uses the document keyword map to identify the keywords from the keyword list that are present in the text. Once the list of keywords present in the text is known, their occurrences in the text are identified and are embedded with hyperlink placeholders. In essence, instances of *keyword* in the source are replaced by,

<a href="javascript:nop" class="cerhyperlink"> *keyword* </a>

which invokes the appropriate display function when selected.

Display of composite links: The related links display is generated when the user clicks on a keyword's hyperlink placeholder. The click event is intercepted by a client side JavaScript function that parses the hyperlink statement and retrieves the actual keyword, sending the keyword to the server

When the request is received at the server, the keyword is retrieved from the request parameters and the metadata repository is used to find all the artifacts related to the keyword. Using these related artifacts, the distinct list of topics to which they belong is identified and links to these topics are generated. For example, if the item has some related image resources then a link to view the images is added to the related links list. Furthermore, the format of these artifacts is also noted. If they are of formats like image or audio then a link to a sample image or audio also is added to the related links list. This sample audio or image is displayed in a new page right on top of the text. This allows the user to view a sample image or listen to a sample audio clip without leaving the text.

The response from the server is received by the client as an XML document object, which is parsed using JavaScript to obtain the related links. The related links are displayed in a tooltip just below the keyword clicked by the user. Cascading style sheets are used to control the look and feel of the tooltip.

6 Discussion and Future Work

This work is one of three major directions we are pursuing to better understand how complex, highly interdisciplinary humanities collections can be designed to enable tight integration resulting in a single, multi-rooted collection. Our focus in the Cervantes music collection has been on leveraging structural information captured as a natural part of the collection building process. Using this information, we are able both to identify link anchors (references to items in the collections) and resources to connect them to. The resulting navigational hypermedia archive enhances the reader's ability to access and interact with the collection as a whole. We would like to expand this approach by more formally investigating the types of structures that can be included in system such as this and the types of automatic linking strategies that each of these structures might support. For example, how might hierarchically structured categories be incorporated? How might that affect the types of interlinkages that can be established? Such an investigation will help us better understand what additional structures are needed to scale this approach to incorporate the full breadth of resources included within the *Cervantes Project* and how it could be generalized to meet the needs of other humanities projects.

In addition to the structural approach to integrating resources presented here, we have also reported on work that uses a formal narrative and thematic taxonomy to provide an integrative framework [2], and also the use of a framework for identifying key features within documents [3]. More work is needed to bring these three directions together to form a unified approach and to understand how each contributes to the larger goal of single, multi-rooted collection.

As we are developing these ideas, we are becoming more aware of the need for a shift in the way we understand the editorial process. Traditional editorial work is focused on the development of a single, centered, completed work that is relatively fixed over time—a published edition. Despite the growing calls to shift from the book as the primary technology for developing scholarly editions to electronic media [16], the resulting editions bear much similarity to their ancestors. In particular, they retain the notion of a "completed" work that is developed to meet narrowly defined research objectives. They are typically created by a single editor or by the highly coordinated efforts of a group of authors working under the guidance of an editorial board. Such editions do not allow for the more complex types of informational needs that are required to support a more broadly defined humanities research agenda, such as that of the *Cervantes Project* [1]. This type of work is open ended and difficult to restrict to a closed set of scholarly perspectives—new research directions continually pop up, often initiated by people outside of the core project members. Its contributors are not the carefully orchestrated cadre of authors one might find in a scholarly encyclopedia (e.g., the *Stanford Encyclopedia of Philosophy* [24]), but rather are individual researchers pursuing their own unique research ideas (and making their own unique contributions). These researchers will often be uncooperative, if for no other reason than their divergent interests (the ethnomusicologist is not likely to be overly concerned about the work of the textual critic), yet their research may contribute significantly to the broader goals of such a digital library. This is not to suggest that traditional approaches are bad or should be abandoned, but rather to propose that we need to creatively explore how to best employ digital technologies to empower humanities research.

References

[1] N. Audenart, R. Furuta, E. Urbina, J. Deng, C. Monroy, R. Saenz, and D. Careaga, "Integrating Diverse Research in a Digital Library Focused on a Single Author," *Proc. 9th European Conf. on Research and Advanced Technology for Digital Libraries*, Vienna, Austria, 2005.

[2] N. Audenart, R. Furuta, E. Urbina, J. Deng, C. Monroy, R. Saenz, and D. Careaga, "Integrating Collections at the *Cervantes Project*," *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital Libraries*, Denver, USA, 2005.

[3] N. Audenaert, R. Furuta, and E. Urbina, "A General Framework for Feature Identification," *Digital Humanities 2006*, to appear.

[4] M. Bernstein, "An Apprentice that Discovers Hypertext Links," *Proceedings of the European Conference on Hypertext*, November 1990, pp. 212-223.

[5] Miguel de Cervantes Saavedra, (1998) *Don Quijote de la Mancha*, Francisco Rico, Director. Barcelona: Biblioteca Clásica, 2 vols; *Don Quixote* (English translation by Edith Grossman) New York: HarperCollins, 2003.

[6] Xin Chen, Dong-ho Kim Kim, N. Nnadi, H. Shah, P. Shrivastava, M. Bieber, Il Im, and Yi-Fang Wu, "Digital Library Service Integration," *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital Libraries*, pp. 384-384, Houston, Texas, 2003.

[7] Cornell University (2005) *Making of America:* http://moa.cit.cornell.edu/moa/ [accessed 8 Sept 2005]

[8] G. Crane, (1998) "The Perseus Project and beyond". *D-Lib Magazine*. http://www.dlib.org/dlib/january98/01crane.html

[9] G. Crane (2000) "Designing Documents to Enhance the Performance of Digital Libraries: Time, Space, People and a Digital Library of London," *D-Lib Magazine* 6 (7/8). http://www.dlib.org/dlib/july00/crane/07crane.html

[10] G. Crane, J.A. Rydberg-Cox, (2000) "New Technology and New Roles: the Need for "corpus editors," *Proc. 5th ACM conference on digital libraries*, San Antonio, TX, pp 252-253

[11] G. Crane, E. David, A. Smith, and C. E. Wulfman, "Building a Hypertextual Digital Library in the Humanities: A Case Study on London," *Proc. 1st ACM/IEEE-CS joint conference on Digital libraries*, pp. 426-434, Roanoke, Virginia, United States, 2001.

[12] Ensemble Durendal. *Por ásperos caminos. Nueva música cervantina.* Ediciones de la Universidad de Castilla-La Mancha, Cuenca, 2005. Text by J. J. Pastor and musical direction by S. Barcellona.

[13] Project Gutenberg Literary Archive Foundation (2005), *Project Gutenberg* http://www.gutenberg.org/. [accessed 9 Sept 2005]

[14] F. Halasz and M. Schwartz. "The Dexter Hypertext Reference Model," *Communications of the ACM*, 37(2), February 1994. pp. 30-39.

[15] B. Ladd, M. Capps, D. Stotts, and R. Furuta. "Multi-head/Multi-tail Mosaic: Adding Parallel Automata Semantics to the Web," *Proc. 4th WWW Conf.*, pp. 422-440, 1995

[16] J. McGann, *The Rationale of Hypertext* : http://www.iath.virginia.edu/public/jjm2f/rationale.html

[17] University of Michigan (2005) Making of America. http://www.hti.umich.edu/m/moagrp/ [accessed 8 Sept 2005]

[18] J. J. Pastor, "Música y literatura: la senda retórica. Hacia una nueva consideración de la música en Cervantes," Doctoral Dissertation, Universidad de Castilla-La Mancha, 2005.

[19] M. S. Patton and D. M. Mimno, "Services for a Customizable Authority Linking Environment," *Proc. 4th ACM/IEEE-CS joint conf. Digital Libraries*, pp. 420-420, Tuscon, AZ, 2004.

[20] H. Plantinga, coord. (2005) Christian Classics Ethereal Library, Calvin College, Grand Rapids, MI. http://www.ccel.org/ [accessed 8 September 2005]

[21] G. Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley, Reading, MA, 1989.

[22] G. Salton, J. Allan, C. Buckley and A. Singhal, "Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts," *Science*, Vol. 264, No. 5164, Jun. 1994, pp. 1421-1426.

[23] F. Shipman, R. Furuta, D. Brenner, C. Chung, and H. Hsieh, "Guided Paths through Web-based Collections: Design, Experiences, and Adaptations," *Journal of the American Society of Information Sciences (JASIS)*, 51(3), March 2000, pp. 260-272.

[24] *Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/, [accessed March 7, 2006].

[25] *Cervantes Project*, E. Urbina, director. Center for the Study of Digital Libraries, Texas A&M University, http://csdl.tamu.edu/cervantes, [accessed Nov 29 2005].

[26] *The Perseus Digital Library*, G. Crane, ed. Tufts University. http://www.perseus.tufts.edu/, [access March 7, 2006].

Building Digital Libraries for Scientific Data: An Exploratory Study of Data Practices in Habitat Ecology

Christine Borgman[1], Jillian C. Wallis[2], and Noel Enyedy[3]

[1] Department of Information Studies
Graduate School of Education & Information Studies, UCLA
borgman@gseis.ucla.edu
[2] Center for Embedded Networked Sensing, UCLA
jwallisi@ucla.edu
[3] Department of Education
Graduate School of Education & Information Studies, UCLA
enyedy@gseis.ucla.edu

Abstract. As data become scientific capital, digital libraries of data become more valuable. To build good tools and services, it is necessary to understand scientists' data practices. We report on an exploratory study of habitat ecologists and other participants in the Center for Embedded Networked Sensing. These scientists are more willing to share data already published than data that they plan to publish, and are more willing to share data from instruments than hand-collected data. Policy issues include responsibility to provide clean and reliable data, concerns for liability and misappropriation of data, ways to handle sensitive data about human subjects arising from technical studies, control of data, and rights of authorship. We address the implications of these findings for tools and architecture in support of digital data libraries.

1 Introduction

The emerging cyberinfrastructure is intended to facilitate distributed, information-intensive, data-intensive, collaborative research [1]. Digital libraries are essential to the cyberinfrastructure effort. As scholarship in all fields becomes more data-intensive and collaborative, the ability to share data becomes ever more essential [2, 3]. Data increasingly are seen as research products in themselves, and as valuable forms of scientific capital [4]. "Big science" fields such as physics, chemistry, and seismology already are experiencing the "data deluge" [5, 6]. Data repositories and associated standards exist for many of these fields, including astronomy, geosciences, seismology, and bioinformatics [7-10]. "Little science" fields such as habitat ecology are facing an impending data deluge as they deploy instrumented methods such as sensor networks. Progress toward repositories and information standards for these fields is much less mature, and the need is becoming urgent.

2 Research Domain

We have a unique opportunity to study scientific data practices and to construct digital library architecture to support the use and reuse of research data. The Center

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 170 – 183, 2006.

for Embedded Networked Sensing (CENS), a National Science Foundation Science and Technology Center based at UCLA, conducts collaborative research among scientists, technologists, and educators. CENS' goals are to develop, and to implement in diverse contexts, innovative wireless sensor networks. CENS' scientists are investigating fundamental properties of these systems, designing and deploying new technologies, and exploring novel scientific and educational applications.

CENS' research crosses four primary scientific areas: habitat ecology, marine microbiology, seismology, and environmental contaminant transport, plus applications in urban settings and in the arts. The research reported here addresses the use of embedded networked sensor technology in biocomplexity and habitat monitoring, supplemented by findings about data sharing from other CENS' areas. In these scientific areas, the goals are to develop robust technologies that will operate in uncontrolled natural settings and in agricultural settings. The science is based on *in situ* monitoring, with the goal of revealing patterns and phenomena that were not previously observable. While the initial framework for CENS was based on autonomous networks, early results revealed the difficulty of specifying field requirements in advance well enough to operate systems remotely. Thus we have moved toward more "human in the loop" models where investigators can adjust monitoring conditions in real time.

3 Background

3.1 Data Digital Libraries and the Data Deluge

Science is a technical practice and a social practice [11]. It is the interaction between technological and social aspects of scientific research that underlies the design challenge. Modern science is distinguished by the extent to which its products rely on the generation, dissemination, and analysis of data. These practices are themselves distinguished by their massive scale and global dispersion. New technologies for data collection are leading to data production at rates that exceed scientists' abilities to analyze, interpret, and draw conclusions. No respite from this data deluge is foreseen; rather, the rate at which data are generated is expected to increase with the advancement of instrumentation [5]. Consequently, scientists urgently require assistance to identify and select data for current use and to preserve and curate data over the long term. Data resources are dispersed globally, due to more international collaboration and distributed access to computing resources for analyzing data. Cyberinfrastructure is expected to provide capabilities to (i) generate scientific data in consistent formats that can be managed with appropriate tools; (ii) identify and extract—from vast, globally distributed repositories—those data that are relevant to their particular projects; (iii) analyze those data using globally distributed computational resources; (iv) generate and disseminate visualizations of the results of such analyses; and (v) preserve and curate data for future reuse. An effective cyberinfrastructure is one that provides distributed communities-- scientific and nonscientific--with persistent access to distributed data and software routinely, transparently, securely, and permanently.

3.2 Data Management Practices

The willingness to use the data of others may be a predictor of willingness to share one's own data. Scholars in fields that replicate experiments or that draw heavily on observational data (e.g., meteorological, astronomical records) appear more likely to contribute data for mutual benefit within their fields. Conversely, scholars in many fields work only with data they have produced. The graph or table that results from analyzing the data may be the essential product of a study. Many scholars assume that the underlying data are not of value beyond that study or that research group. Heads of small labs often have difficulty reconstructing datasets or analyses done by prior lab members, as each person used his or her own methods of data capture and analysis. Local description methods are common in fields such as environmental studies where data types and variables may vary widely by study [12, 13].

The degree of instrumentation of data collection also appears to be a factor in data sharing. Sharing expensive equipment is among the main drivers for collaboration, especially in fields such as physics and chemistry. In these cases, collaboration, instrumentation, and data sharing are likely to be correlated. The relationship between instrumentation and data sharing may be more general, however. A small but detailed study conducted at one research university found that scholars whose data collection and analysis were most automated were the most likely to share raw data and analyses; these also tended to be the larger research groups. When data production was automated but other preparation was labor-intensive, scholars were less likely to share data. Those whose data collection and analysis were the least automated and most labor-intensive were most likely to guard their data. These behaviors held across disciplines; they were not specific to science [14].

3.3 Habitat Ecology Data and Practices

The study of biodiversity and ecosystems is a complex and interdisciplinary domain [15]. The mechanisms used to collect and store biological data are almost as varied as the natural world those data document. Over the last thirty years, data management systems for ecological research have evolved out of large projects such as the *International Biological Program* (IBP; established in 1964 by the International Council of Scientific Unions), the *Man and the Biosphere* program (MAB; established in 1971 by the United Nations), and the U.S. *Long-Term Ecological Research* program (LTER; established in 1980 by the National Science Foundation) [16]. These systems need to support multiple data types (numerical measurements, text, images, sound and video), and to interact with other systems that manage geographical, meteorological, geological, chemical, and physical data. Currently one of the biggest challenges to the development of effective data management systems in habitat ecology is the "datadiversity" that accompanies biodiversity [17].

The Knowledge Network for Biocomplexity (KNB) [http://knb.ecoinformatics.org/], an NSF-funded project whose first products became available in 2001, is a significant development for data management in habitat ecology. KNB tools include a data management system for ecologists, based on the Morpho client software and Metacat server software, and a standard format for the documentation of ecological data—the Ecological Metadata Language (EML). SensorML is an equally important development for sensor data [18].

4 Research Problem

While practices associated with scholarly publication vary widely between fields, the resulting journal articles, papers, reports, and books can be described consistently with bibliographic metadata. Data are far more problematic. Disciplines vary not only in their choice of research methods and instruments, but the data gathered may vary in form and structure by individual scholar and by individual experiment or study. Multidisciplinary collaboration, which is among the great promises of cyberinfrastructure, will depend heavily on the ability to share data within and between fields. However, very little is yet known about practices associated with the collection, management, and sharing of data. Despite these limitations, immediate needs exist to construct systems to capture and manage scientific data for local and shared use. These systems need to be based on an understanding of the practices and requirements of scientists if they are to be useful and to be used.

Habitat ecology is a "small science," characterized by small research teams and local projects. Aggregating research results from multiple projects and multiple sites has the potential to advance the environmental sciences significantly. The choice of research problems and methods in environmental research were greatly influenced by the introduction of remote sensing (satellite) technology in the 1980s and 1990s [19]. Thus one of our research concerns is how habitat ecology may evolve with the use of embedded networked sensing. These scientists are deploying dense sensor networks in field locations to pursue research on topics such as plant growth, bird behavior, and micrometeorological variations.

Our research questions address the initial stages of the data life cycle in which data are captured, and subsequent stages in which the data are cleaned, analyzed, published, curated, and made accessible. The questions can be categorized as follows:

- **Data characteristics:** What data are being generated? To whom are these data? To whom are these data useful?
- **Data sharing:** When will scientists share data? With whom will they share data? What are the criteria for sharing? Who can authorize sharing?
- **Data policy:** What are fair policies for providing access to these data? What controls, embargoes, usage constraints, or other limitations are needed to assure fairness of access and use? What data publication models are appropriate?
- **Data architecture:** What data tools are needed at the time of research design? What tools are needed for data collection and acquisition? What tools are needed for data analysis? What tools are needed for publishing data? What data models do the scientists who generate the data need? What data models do others need to use the data?

5 Research Method

5.1 Data Sources

Our goal is to understand data practices and functional requirements for CENS ecology and environmental engineering researchers with respect to architecture and policy, and to identify where architecture meets policy. The results reported here are drawn from multiple sources over a three-year period (2002-2005). In the first year (2002-2003), we sat in on team meetings across CENS scientific activities and we inventoried data standards for each area [20]. In year 2 (2003-4), we interviewed individual scientists and teams and continued to inventory metadata standards. We used the results of the first two years to design an ethnographic study of habitat biologists, conducted in year 3 (2004-05). In the current year (2005-6), we are interviewing individual members of habitat ecology research teams, including scientists, their technology research partners in computer science and engineering, and graduate students, postdoctoral fellows, and research staff.

5.2 Process

The ethnographic work from the first three years of the study (interviewing teams and individuals, participating in working groups, etc.) is documented in notes, internal memoranda, and a white paper [20]. We did not audiotape or videotape these meetings to avoid interfering with the local activities. Knowledge from this part of the research was used to identify data standards relevant to the research areas. We shared our results with individuals and teams to get feedback on the relevance of these standards. We also constructed prototypes of data analysis and management tools as components of the educational aspects of our research [21]. Thus we are conducting iterative research, design, and development for data management tools in CENS.

5.3 Participants

Our population at CENS is comprised of about 70 faculty and other researchers, a varying number of post-doctoral researchers, and many student researchers. About 50 scientists, computer scientists, engineering faculty, and their graduate students, post-doctoral fellows, and research staff are working in the area of habitat ecology. The data reported here are drawn primarily from in-depth interviews of two participants, each two to three hours over two to three sessions. The direct quotes are from these interviews. Results from one-hour interviews with two other scientists and from a large group meeting (about 20 people) to discuss data sharing policy also are reported here. These results are informed by interviews, team meetings, and other background research conducted in earlier stages of our data management studies.

5.4 Analysis

We used the results of interviews and documentary analyses in the first two years of our research to design the ethnographic study. This study used the methods of grounded theory [22]. The interview questions are based on Activity Theory [23-25], which analyzes communities and their evolution as “activity systems.” Activity systems are defined by the shared purposes that organize a community and by the ways in which joint activity to achieve these purposes is mediated by shared tools, rules for behavior, and divisions of labor. When analyzing how activity systems change and develop, the focus is on contradictions that occur within the system or as a result of the system interacting with other activity systems. These contradictions are analyzed as the engine for organizational change.

Based on this theoretical framework, we developed interview questions about participants' motives, understanding of their community's motives, tools used in daily work, ways they divided labor, power relations within their community, and rules and norms for the community. Interviews were then fully transcribed. In the initial phase of analysis we looked at the first interviews with participants for emergent themes. The analysis progressed iteratively. Subsequent interviews were analyzed with an eye towards testing and further refining the themes identified in the initial coding. With each refinement, the remaining corpus was searched for confirming or contradictory evidence. At this stage, however, the work is still preliminary. As such, no formal coding schemes were developed that were systematically applied to the entire corpus. Rather, what we present below are the emergent themes and representative illustrations in the participants' own words.

6 Results

In the first two years of study, we learned that CENS' habitat biologists perceived a lack of established standards for managing sensor data, specifically those that support the sharing of data among colleagues. They are eager to work with us because they need tools to capture, manage, and share sensor data more effectively, efficiently, and easily. They are committed to participation of developing domain data repositories and standards such as the Knowledge Network for Biocomplexity and Ecological Metadata Language, but are not yet implementing them. A good starting point for exploring the metadata requirements of this community is to assess scientists' experiences with implementing these tools and standards, and to evaluate those tools' utility. The results are organized by the research questions outlined above: Data characteristics, data sharing, data policy, and data architecture.

6.1 Data Characteristics

Our interview questions explored what data are being generated, to whom are these data, and to whom are they useful. CENS is a collaboration between technologists and scientists, thus the technologies are being evaluated and field tested concurrently with the scientific data collection. The scientists interviewed reported unreliability of the sensors. In the early stages, one researcher found that he was losing about 25% of every transmission from every sensor, for example. While the sensors were sending data every five minutes, they only produced usable data every 10 or 15 minutes. Thus they could not always trust what they were getting from the instruments:

> *I'm highly suspicious [of automated data collection]. I mean it works, but then sometimes it doesn't, and then sometimes weird things happen. You get a glitch, and then you start getting the same value twice or something. … when you do averages, it's all funny.*

These researchers have a story in mind when they design a field experiment. They also know about how much data they will need to support that story in a published paper in their discipline. One of our scientists explicitly told us that

> *… to tell that story I'm going to need an average of five figures and a table.*

He sketches out mock figures on paper as part of his research design. We were particularly intrigued by his notion of "least publishable unit" – in this case, five figures and a table. However, he also told us that he tends to collect more data so that he can elaborate on his story:

> *I collect another data set just to round it out rather than [picking] an interesting sub-phenomenon of another sub-phenomenon. That's boring.*

Thus his research design is based on the amount of data he needs to tell a story of this scope and form. The publication is the product of the study, rather than the data per se.

6.2 Data Sharing

We collected some useful commentary on issues of what data scientists will share, when, with whom, and with what criteria. Within this small sample, scientists generally are more willing to share data that already have been published and less willing to share data that they plan to publish. The latter type of data represent claims for their research.

> *Sharing data- if it's already published? It's your data, no problem. I can give that to [you]. If it's something I'm working on to get published or somebody else is working on to get published, or if they want to publish the paper together, it gets a little bit funnier.*

For this scientist, willingness to share also is influenced by the effort required to collect the data. His hand-collected data are more precious than his instrument-generated data:

> *… if you walk out into a swamp .. out in this wacky eel grass, and marsh along with your hip-waders and [are] attacked by alligators ..and then you do it again and again and again... I don't [want to] share that right away. I [want to] analyze it because I feel like it's mine.*

The above dataset was seen as "hard won." When they do share experimental data with collaborators, they feel an ethical and scientific responsibility to clean the dataset sufficiently that it has scientific value. Raw data is meaningless to others. Unless the data are useful and relevant, they would be "just taking up space and nobody's going to be able to use [them]."

If they feel they are forced to share data they will, knowing that it may not be of much use to others. One scientist told us if someone wants his data, he or she can have it in the raw form. His Excel spreadsheets are cryptic and exist in multiple versions, representing each transformation.

Conversely, we found less evidence of proprietary ownership over reference data that provides context, but is not specifically relevant to their research questions. One scientist gave the example of measuring the density of shade cloth for a field experiment. Much work went into determining the amount of shade a particular type of cloth provided in a 24-hour period. He was happy to save other people the effort of reconstructing that number. Similarly, our subjects usually are willing to share software and other tools.

Several of the researchers interviewed did not think their data would be of much use to other researchers. Conversely, some said the data they are collecting already is being shared between themselves and statisticians, engineers, and computer scientists, all with different purposes for the data. One of our subjects recognized the possible uses of his data on water contaminants in a river confluence for such diverse fields as fluid mechanics, public health, ichthyology, and agriculture.

6.3 Data Policy

We encountered a particularly enlightening scenario in which questions arose of whether the data from an instrument belonged to the designers of the instrument or the designers of the experiment. The team member (engineering faculty) who designed and installed the instrument had plans for using the data from the instrument but did not implement those plans. After several years of data production, one of the scientists found the data useful for his own research, and asked the head of the research site for permission to use the data in a publication. Given that no other claims were being made on the data, and that these data were being posted openly on the website of the research site, permission was granted. After some investment in cleaning and analysis, the data looked promising for publication, so the scientist and site director invited the instrument designer to participate in the publication. However, the designer objected strongly on the grounds that they were his data because he had deployed the instrument. The situation was complicated further by the existence of a pending grant proposal involving this instrument by the same designer. We learned later that the situation was resolved only when the pending grant was not awarded, relieving some of the proprietary tension. The scientist we interviewed commented that this was the first real intellectual property issue over data that he had encountered.

In the above case, the technology people are faculty partners in the research. Yet they view the status of the data and the control over it rather differently. Authorship credit on publications is a common issue in research. In cases where instrumentation is essential to the data collection, questions sometimes arise as to whether technical support people should be authors. In another interview, the scientist who provided the above example commented that

> *... tech-support people might get an acknowledgement ... but they're not co-authors on a scientific paper.*

The two situations described here are distinctly different. In the former, the technologist was a researcher who had deployed the instrument, and all agreed that he was entitled to some form of authorship credit. The issue appeared to be about who had priority over the data in determining what should be published, when, and by whom. In the latter situation, a scientist is referring to people who assist with equipment but are not themselves researchers. However, situations may arise where the distinction between technical research and technical assistance is not clear.

In the group meeting (about 20 CENS faculty, students, and research staff) to discuss the ethics and policy of data sharing, several interesting issues arose. One frequent question was about the condition of data to be shared. The group generally agreed that those generating the data had responsibility to assure that the data were reliable and were verified before sharing or posting. Most participants were aware of NSF rules for sharing the data from funded research. At the same time, they were concerned about premature release prior to publication, and whether any sort of liability disclaimers or rights disclaimers (e.g., Creative Commons licenses for attribution and non-commercial reuse) should be applied.

Several of the meeting participants were involved in a small project involving cameras triggered by sensors. The purpose of the project was to test the sensors, but capturing identifiable data on individuals might be unavoidable. They were very concerned about privacy and security issues with these data. Because the study was not about human subjects, they had not sought human subjects review. Several people analogized the situation to webcams on university campuses. Images of people are streaming to public websites without the knowledge or permission of those involved. They also discussed technical solutions such as anonymizing faces captured by the cameras.

6.4 Data Architecture

A longer term goal of our research is to design tools to support data acquisition, management, and archiving. A number of questions addressed the use of data and tools at each stage in the life cycle.

6.4.1 Research Design and Hypothesis Testing

Field research in habitat ecology begins with identifying a research site in which the phenomena of interest can be studied. Before scientists begin setting up sensors or deciding which extant sensors might produce data of interest, they would like a map of the research site that includes the location of sensors and the types of data that each sensor could produce. For example, this scientist would like a map of the area and a table of the data from each set of sensors:

> *I want a table that I can skim really easily and say Okay of these ten stations how many of them have temperature data available? I don't want to look around on a map and have to click on each link*

Scientists spend much time on activities such as sensor placement and development prior to fieldwork. They test equipment and sample the quality of observations from the sensors before they start doing any real science. Thus tools for this exploratory phase are desirable.

6.4.2 Data Collection and Acquisition

Due largely to the relative immaturity of the sensor technology, several of our science subjects were suspicious of automated data collection. They expressed reluctance to take data streams straight from the sensors without good tools to assess the cleanliness of the data. They want simple and transparent methods to find potential gaps in the data; also desirable are tools to identify when values are duplicated or missing, when sensors are failing, and other anomalous situations.

Some also expressed the need to annotate the data in the field, which would provide essential context that cannot be anticipated in data models. For example, data collection tools can have default menus for the available data elements generated by any particular sensor. What they cannot do in advance is predict how scientists will modify the instruments or the field conditions. One scientist mentioned moving his equipment to a different location to compare temperatures. Documenting which data were collected at which location and when is essential to later interpretation. Another good example is distinguishing between common instruments with known characteristics and one that was hand-made. Another scientist might use the same off-the-shelf instrument for another experiment, enabling easy comparisons. If the instrument were unique, the results may be much more difficult to interpret. The scientist might calibrate the two instruments and find they could be used interchangeably, but the future user of the data needs to know what instruments were used and how. The example here is

> *air temperature [from] a little therma-couple that I stuck one foot off the ground with a little aluminum shield that I made. Someone else might use [a common purchased instrument] interchangeably, but they should know that one of them was my home-built little therma-couple thing.*

6.4.3 Data Analysis

Cleaning. The two scientists in the ethnographic study would extract data from the sensor network database to perform any correlations. They prefer to use graphing programs with which they are most familiar. One scientist was disinterested in the offer of graphing tools or visualizations, because he was reluctant to trust other people's graphs. He wants his own graphs so that he can make his own correlations. He also wants the data in a form that he can import into his preferred analysis programs. Data are cleaned with respect to specific research questions. If data are extraneous to a current paper, they may not be cleaned or analyzed. Thus the datasets resulting from a field project are not necessarily complete.

Version Control. Even when scientists are willing to share archived data, those data may be poorly labeled, rendering them difficult to use. Multiple versions may exist of the same data set, resulting from different cleaning or analysis processes. Most of these processes are not recorded, making the data even less accessible to the potential consumer. Multiple versions of datasets complicate retrieval for the scientists who created them and for future researchers who may want to use them. One of our researchers acknowledged that each person on the team created his or her own Excel spreadsheets, and the only access to the data of their teammates was to ask for the spreadsheets and explanations of their contents. This scientist worried that when any of her team members left the project, their data essentially would be lost.

Tools. In these interviews, and in prior interviews and meetings over the last several years, we often found that scientists prefer viewing data in tables, especially Microsoft Excel spreadsheets. One scientist in the ethnographic study offered a detailed explanation. Columns and tables enable him to identify holes in data and to determine how to clean them. Graphs and plots show different types of data inconsistencies, such as identifying dead sensors or graphics in the wrong time scale. Graphs are also personal, because scientists reduce data to test their own hypotheses.

These scientists do not appear to trust transformations made by others; they are more likely to apply their own tools and methods to the original data. The scientist noted above would trust fellow biologists to clean the data:

> *I want some radiation measurements ... you can do energy budget calculations that … have to be cleaned up by a biologist in order to get incorporated into the data set.. any old biologist can come by and start doing correlations, because they know what the data is...*

7 Discussion and Conclusions

While the number of interviews reported in this study is small, the results are based on four years of work with these scientists and technology researchers. Collaborative work can be much slower than solo work due to the effort involved in learning a common language and in finding common ground in research interests [26]. The investments pay off in new insights and new approaches. CENS collaborations between scientists and technologists did not lead to new forms of science as quickly as expected. In its fourth year of existence, the Center is now deploying networked sensor technologies for multiple scientific studies. The promised payoffs are beginning to accelerate. Many of the problems with data cleaning and sensor reliability are due to the immature stage of the instruments and networks. As these scientists become more experienced with these technologies, they are likely to experiment with more instruments and configurations, so the data cleaning and calibration issues will not go away. Experience also is likely to make them more discriminating consumers of the technology, making yet greater demands on the technology researchers.

Scholarly publications have been the product of science for several centuries. Viewing data as a product per se is a relatively new idea. The scientists that we interviewed for this study continue to focus on the paper as their primary product, designing their experiments accordingly. Whether the data will become a direct product of their research to be reused and shared is one of our continuing research questions.

Our tentative findings about data sharing are consonant with other research: In this small sample, our scientists are more willing to share data already published than data that they plan to publish [27]. One scientist was explicit about guarding hand-collected data more closely than data from sensor networks [14]. Our subjects expressed responsibility to assure that any shared data are documented sufficiently to be interpreted correctly [27]. If required to share data they will, knowing that raw data are of little value to others without sufficient cleaning and documentation. However, given a choice, most prefer to exploit their research data fully before releasing them to others.

A number of interesting policy issues arose that we are now studying in much more depth. The Center has made a commitment to share its data with the community and is seeking ways to do so. Members want to provide clean and reliable data, but are understandably concerned about liability and misappropriation of data. Among the questions to address are how to handle sensitive data about human subjects that arises from technical studies, who controls data from a project, and who has first rights to authorship. These are common issues in collaboration, especially at the boundaries between fields. The boundary between life sciences, such as habitat biology, and technology may be even more complex. Not only do differences in data practices between these domains arise, but the distinction between technical research and technical assistance may not always be clear.

These findings have some promising implications for architecture and tools for data management in habitat ecology and perhaps to other field research disciplines. One is the need for tools to explore the research site and the availability of extant data sources. These are especially valuable in the early design stages of an experiment. Visualization tools in the field may be less helpful than expected, as these scientists want to get the data into their own, familiar tools. Quick prototyping of data sources in the field is an essential requirement, and one already recognized in CENS' "human in the loop" architecture. These scientists wish to add new instruments and new details about data and instruments in the field on an ad hoc basis. They invent new tools as needed, using aluminum foil, cloth, tape, and other available equipment. They want data analysis in the field so that they can adjust experiments in real time.

Most of the above requirements suggest hand-crafted tools and structures for this research community. The longer term goal, however, is to build generalizable, scalable tools that facilitate sharing and curation of scientific data. We will continue to work with habitat biologists and other CENS scientists to find a balance between local and global requirements for tools and architecture. While the study reported here relies on a small dataset and is exploratory in nature, it identifies a number of important questions for the design of cyberinfrastructure for science. Research to pursue these questions in more depth is currently under way.

Acknowledgements

CENS is funded by National Science Foundation Cooperative Agreement #CCR-0120778, Deborah L. Estrin, UCLA, Principal Investigator. CENSEI, under which much of this research was conducted, is funded by National Science Foundation grant #ESI-0352572, William A. Sandoval, Principal Investigator and Christine L. Borgman, co-Principal Investigator. Jonathan Furner and Stasa Milojevic of the Department of Information Studies provided valuable material to the fieldwork on which this study is based. Robin Conley and Jeffrey Good, UCLA graduate students, assisted in interviewing and data transcription.

References

1. Atkins, D.E., et al., *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon panel on Cyberinfrastructure*. 2003, National Science Foundation: Washington, D.C.
2. Unsworth, J., et al. *Draft Report of the American Council of Learned Societies' Commission on Cyberinfrastructure for Humanities and Social Sciences*. Last visited 5 November 2005 http://www.acls.org/cyberinfrastructure/acls-ci-public.pdf.
3. Berman, F. and H. Brady. *Final Report: NSF SBE-CISE Workshop on Cyberinfrastructure and the Social Sciences*. Last visited 18 May 2005 http://vis.sdsc.edu/sbe/reports/SBE-CISE-FINAL.pdf.

4. Schroder, P., *Digital Research Data as Floating Capital of the Global Science System*, in *Promise and Practice in Data Sharing*, P. Wouters and P. Schroder, Editors. 2003, NIWI-KNAW: Amsterdam. p. 7-12.
5. Hey, T. and A. Trefethen, *The Data Deluge: An e-Science Perspective*, in *Grid Computing – Making the Global Infrastructure a Reality*. 2003, Wiley.
6. Hey, T. and A. Trefethen, *Cyberinfrastructure and e-Science.* Science, 2005. **308**: p. 818-821.
7. *International Virtual Observatory Alliance*. Last visited 2 March 2005 http://www.ivoa.net/.
8. *Incorporated Research Institutions for Seismology*. Last visited 25 November 2004 http://www.iris.edu.
9. *Biomedical Informatics Research Network*. Last visited 19 March 2005 http://www.nbirn.net/.
10. *GEON*. Last visited 19 March 2005 http://www.geongrid.org/.
11. Star, S.L., *The politics of formal representations: Wizards, gurus and organizational complexity*, in *Ecologies of Knowledge: Work and Politics in Science and Technology*, S.L. Star, Editor. 1995, State University of New York Press: Albany, NY.
12. Estrin, D., W.K. Michener, and G. Bonito, *Environmental cyberinfrastructure needs for distributed sensor networks: A report from a National Science Foundation sponsored workshop*. 2003, Scripps Institute of Oceanography.
13. Zimmerman, A., *New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data.* Science, Technology, & Human Values, under review.
14. Pritchard, S.M., L. Carver, and S. Anand, *Collaboration for knowledge management and campus informatics*. 2004, University of California, Santa Barbara: Santa Barbara, CA. Retrieved from http://www.library.ucsb.edu/informatics/informatics/documents/UCSB_Campus_Informatics_Project_Report.pdf on 14 November 2005.
15. Schnase, J.L., et al. *Building the next generation biological information infrastructure*. in *Proceedings of the National Academy of Sciences Forum on Nature and Human Society: The Quest for a Sustainable World*. 1997. Washington, DC: National Academy Press.
16. Michener, W.K. and J.W. Brunt, eds. *Ecological Data: Design, Management and Processing*. 2000, Blackwell Science: Oxford.
17. Bowker, G.C., *Biodiversity datadiversity.* Social Studies of Science, 2000. **30**(5): p. 643-683.
18. Brown, C., *Lineage metadata standard for land parcels in colonial states.* GIS/LIS '95 Annual Conference and Exposition. American Soc. Photogrammetry & Remote Sensing & American Congress on Surveying & Mapping. Bethesda, MD, USA., 1995. Part 1, Vol 1: p. 121-130.
19. Kwa, C., *Local ecologies and global science: Discourses and strategies of the International Geospher-Biosphere Programme.* Social Studies of Science, 2005. **35**(6): p. 923-950.
20. Shankar, K., *Scientific data archiving: the state of the art in information, data, and metadata management.* 2003.
21. Sandoval, W.A. and B.J. Reiser, *Explanation-driven inquiry: Integrating conceptual and epistemic supports for science inquiry.* Science Education, 2003. **87**: p. 1-29.
22. Glaser, B.G. and A.L. Strauss, *The discovery of grounded theory: Strategies for qualitative research*. 1967, Chicago: Aldine Publishing Co.
23. Engeström, Y., *Activity theory and individual and social transformation*, in *Perspectives on activity theory.* 1999, New York: Cambridge University Press: p. 19-38.

24. Engeström, Y., *Learning by Expanding: An activity-theoretical approach to developmental research*. 1987, Helsinki: Orienta-Konsultit.
25. Cole, M. and Y. Engeström, eds. *A Cultural-historical Approach to Distributed Cognition.* 1993, New York: Cambridge University Press.
26. Cummings, J.N. and S. Kiesler, *Collaborative research across disciplinary and organizational boundaries.* Social Studies of Science, 2005. 35(5): p. 703-722.
27. Arzberger, P., et al., *An International Framework to Promote Access to Data.* Science, 2004. 303(5665): p. 1777-1778.

Designing Digital Library Resources for Users in Sparse, Unbounded Social Networks

Richard Butterworth

Interaction Design Centre, School of Computing Science,
Middlesex University, London, UK NW4 4BT.
r.j.butterworth@mdx.ac.uk
http://www.cs.mdx.ac.uk/staffpages/richardb/

Abstract. Most digital library projects reported in the literature build resources for dense, bounded user groups, such as students or research groups in tertiary education. Having such highly interrelated and well defined user groups allows for digital library developers to use existing design methods to gather and implement requirements from those groups. This paper, however, looks at situations where digital library resources are aimed at much more sparse, ill defined networks of users. We report on a project which explicitly set out to 'broaden access' to tertiary education library resources to users not in higher education. In particular we discuss the problem of gathering *á priori* user requirements when by definition, we did not know who the users would be, we look at how disintermediation plays an even stronger negative role for sparse groups, and how we designed a system to replicate an intermediation role.

1 Introduction

If one were to consider a 'typical' digital library (DL) project reported in the literature one is likely to think of a digital library resource based on university library holdings and aimed at students or academic researchers (eg. [1, 2, 3]). The user groups in this case:

are well defined — it is possible to tell who is and who is not a potential user of the DL system,

have well defined needs and tasks — it is possible to tell what they want to use the system for,

are co-located or easily accessible — it is not expensive to question them to gather their requirements,

are homogeneous — their requirements are broadly similar; if you have a user population of one hundred undergraduate students doing the same course, interviewing, say, ten of them is likely to give an adequate picture of the group as a whole, and

are highly interrelated — the individuals in the groups tend to be closely related in their work, research or studies.

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 184–195, 2006.

(Note that although we assert that it is *possible* to discover the boundaries and needs of such groups, we do not suggest it is particularly *easy*.)

In network analysis [4] such collections of users are called ‘dense, bounded groups’. This work, however, contends that many users of both traditional and digital libraries are not sufficiently accurately modeled by dense bounded groups, particularly when looking at user networks outside tertiary education. If dense, bounded groups are at one end of a continuum, then at the other end are ‘sparse, unbounded networks’. It is these networks of users we look at in this work, and report how we set to out design better digital library resources for them.

1.1 Dense, Bounded Groups and Sparse, Unbounded Networks

Network analysis (eg. [5, 6]) is a branch of social science that looks at the structure of social networks, in particular analysing the relationships between the actors in networks. The domain of social networks looked at by network analysis is very broad: from analyses of markets to the structure of riots, but it is the work of Wellman [4] that applied network analysis to the field of IT, by characterising the different social networks that can be mediated online, and it is Wellman’s description of the difference between dense, bounded groups and sparse, unbounded networks that we base our work on.

Dense, bounded groups[1] are characterised by networks that have well defined boundaries and a high degree of interrelationship between the actors in the network. Typically the starting point for an analysis of a dense bounded group is the definition of the boundaries of the group: this implies who is or is not inside the group, and analysis can proceed on the group members.

In contrast sparse, unbounded networks are characterised by relationships that cross formal boundaries. For example, a formal boundary may be organisational: there is a clear line around who does and does not work for a particular organisation. Networks that cross these boundaries may be friendship networks, or networks of common interests. Because by definition we cannot start an analysis by defining the boundaries of an unbounded group, analyses of sparse, unbounded groups start by looking at the relationships of one or two individuals and then traces their relationships outwards. If boundaries are discovered, then they *emerge as a consequence* of the analysis, not as in bounded groups where they are the starting point for the analyses.

Wellman describes one of the characteristics that differentiate sparse networks and bounded groups is that the relationships in sparse networks ‘tend to ramify out in many directions like an expanding spider’s web’ whereas the relationships in dense groups ‘curl back on themselves into a densely knit tangle’ [4, page 180]. Note that the difference between the two types of networks is defined both internally by the characteristics of the relationships in the network, and externally by the way that they are analysed.

[1] Note that the term ‘group’ has a specialised meaning in network analysis: a dense, bounded network is referred to as a ’group’. In this paper we shall adhere to this specialised terminology.

1.2 Is the Difference Important?

This paper gives examples of digital library users that are much better characterised as sparse, unbounded networks. We argue that standard development methodologies are not ideal for designing systems with unbounded user networks, and that intermediation is critical for sparse networks. But before moving on we need to address the question: how much does it matter? Even though standard development methodologies and disintermediated DL models are based on the assumption that user groups are dense and bounded, can we still use those methodologies and models to develop good DL resources?

Evidently the answer is yes: good DL resources have been developed using standard methodologies. However we would argue that the risk of failure is higher because of this mismatch between the assumed and actual characteristics of the user populations. Furthermore as we argued above much reported DL work has been developing resources for tertiary education users, where user groups are generally dense and bounded. However, once we move outside the tertiary education domain the evidence for successful DL projects becomes weaker (see [7]). There are many difference reasons for this lack of success, but we suggest that user networks outside tertiary education being much sparser and unbounded, is a contributing factor.

2 The Accessing Our Archival and Manuscript Heritage Project

The Accessing our Archival and Manuscript Heritage (AAMH) project was a fourteen month project undertaken at Senate House Library, University of London which aimed to develop online resources to encourage and assist life-long learners to use the materials held in University of London libraries and archives[2]. The project was particularly aimed at opening up access to the libraries' special collections and archives. It was felt that these collections held much material that would be of benefit to users outside tertiary education.

The explicit aim of the project was the broaden access to library resources. Precisely how this broadening of access was to be facilitated was not explicitly addressed in the early project proposal. It was up to the project staff to decide (for example) whether directly surrogating library resources by digitising materials or by the more indirect route of surrogating library services would best fulfil the remit of the project.

Taken to its furthest implications 'broadening access' means that we *could not* know beforehand who the users of the proposed system would be. We would have to build it and see who came; we could not perform *á priori* user requirements gathering. But given that non-existant, incomplete, changeable or otherwise ill defined requirements are often quoted [8] as the main culprit in project failure

[2] The University of London is a federated university, consisting of several colleges, and academic institutions. Many of the constituent colleges and institutions have their own libraries and archives.

this seemed to be a recipe for potential disaster. A search of requirements and software engineering texts for methodologies that helped us gather user requirements when we did not know who the users were was, unsurprisingly, fruitless. However some requirements gathering methodologies showed more applicability to our situation than others, and in the next section we outline the methodology we used and review how it worked in practice.

In order to get started we had to devise some assumptions about who our potential user base would be. After liaison with several of the university archivists it became clear that the resources in their archives was most of use to outside of the usual students and academics were 'amateur' family and local history researchers, who we refer to collectively as 'personal history researchers'. Even though deciding to initially limit ourselves to personal history researchers set some sort of bounds on our user population, this user population was still fairly unbounded and sparse.

Compare the characteristics of these users to that of the 'typically' reported user population set out at the beginning of this paper. They:

are only very loosely bounded — an interest in personal history hardly constitutes a limiting boundary: who is *not* interested in their family history?

do not have well defined needs and tasks — there are a multitude of ways of tracing your family tree, particularly once researchers have moved beyond the basic census and birth, marriage and death registers.

are not easily accessible and co-located —personal history researchers are quite happy to work on their own: how does one find and contact these researchers to analyse their needs?

are extremely heterogeneous — in talking to several members of local and family history groups we encountered researchers with an enormous range of skills, from researchers who had no training in research skills to a retired history professor.

are only weakly interrelated — many researchers we contacted were members of local or family history groups, but these met occasionally (typically monthly) and in most cases this was the only contact they had with other similar researchers.

All this adds up to a sparse, unbounded user population.

Note that the AAMH project was action research. The main outcome of the project was a working, useful DL resource: we did not explicitly set out to develop for sparse, unbounded networks of users. That the users we were looking at shared characteristics with models described in social science literature emerged as a consequence of the design work we were doing. The work described below is therefore a largely *post hoc* rationalisation, looking at the work we did through the lens of network analysis.

3 Iterative Requirements Gathering and Implementation

The software and requirements engineering literature (eg [9, 10]) was surveyed, but we could not find a methodology that suited our needs. It is clear that most software development and requirements gathering processes are based on the assumption that the projected user groups for the system to be developed are bounded groups. Most requirements engineering methods, advise that the developers should first define who the users are, and then carefully and explicitly gather and analyse their requirements. Clearly the process of 'defining who the users are' is about setting boundaries on the group, and gathering requirements takes place within that defined group: this is an exact parallel to the approach to analysing bounded groups set out in the introduction.

Furthermore DL systems are highly interactive user driven systems, and therefore to ensure usability and usefulness there are strong arguments [11, 12] that an iterative design process is needed. Simply put, an iterative design process gathers requirements from a user group, rapidly prototypes an implementation that hopefully meets those requirements, then evaluates the implementation with the users. Evaluation will suggest changes to the prototype or to the requirements, and the process iterates taking these changes into account. The idea is that from an approximately acceptable starting prototype an increasingly acceptable implementation is developed.

The spiral model [13] is about iteratively developing prototypes whereas the star model [11] shows that the requirements should also be included in the process. The star model also argues that developers do not work in a linear way from requirements to implementation at all: they may start with a prototype, and then work 'backwards' so that the requirements for the prototype emerge. The key point is that whenever any artefact (requirements statement, prototype, etc) is proposed it should be evaluated before moving on to develop further design artefacts.

However the unbounded nature of our users posed problems for these iterative process. Recall that the boundaries of an unbounded network emerge (if at all) as a consequence of analysing the network, in other words, we would have to do a lot of analysis in order to delimit who the users actually are, *before* we could embark on the sort of iterative design process described above. In a time limited project like AAMH this was not practical: once we had an analysis of our user population that was good enough to use in the design, we were likely to have run out of time to develop anything. Therefore a different approach was needed such that the analysis of the unbounded user network took place at the same time as the DL resources were being developed.

3.1 'Early Phase' Requirements Engineering

However, 'early phase' requirements engineering [14] offered promise. The key principle in standard requirements engineering is that requirements state *what* a system should do, as opposed to *how* the system should do it. This should promote a clearer understanding of the system among the designers, who are liable to lose track of what a system should be doing among the messy details of how it does it. 'Early phase' requirements engineering goes one step further, not only describing what a system should do, but *why* it should do it. These 'why' statements should promote a clearer understanding not only of the projected system, but the context (organisational, environmental, user, etc) in which it sits.

Early phase requirements engineering looked valuable because it is the why statements — the understanding of the users — which would hopefully emerge as the project progressed. We therefore proposed to use informal early phase requirements engineering in an iterative manner to develop our DL resources.

3.2 Our Proposed Design Process

Using early phase terminology there are three groups of design artefacts: why statements which describe context and assumptions, what statements which describe requirements, and how statements which describe implementations. As described above the spiral model is about iteratively refining how statements, and the star model iteratively refines what and how statements. The innovation of our design process is that it includes why, what and how statements in the process. In effect our design process was an augmentation of the star model, where contextual assumptions are also treated as design artefacts.

A likely consequence of such a process would be that we would not get a 'neat' incremental improving of the prototype: changes in the why statements were likely to result in very dramatic changes to the prototype. Such largely changes are probably unavoidable, but the important point is that the project expects them, and leaves enough slack in the project schedule to deal with them.

In our case the design process would start with a set of educated guesses about who the potential users might be and a broad description of their characteristics (the why statements), what their needs would be (the what statements) and a rapid prototype of a DL system that met those needs (the how statements). We would then evaluate these three sets of statements with potential users, change them according to the evaluation, and then iterate.

3.3 The Design Process in Action

Space precludes a detailed description of how this design process worked in action on the AAMH project (see [15, section 5] for a more detailed account), but we include a sketch here to demonstrate the value that this design process added to the project.

First iteration. Our initial 'why' statement proposed that our potential users would be people interested in using library archive materials in their research. We further proposed a model of the four processes they would engage in to use archive material. We suggested that they would:

- propose research questions,
- identify archival collections that would help answer those questions,
- search for materials in those collections, and
- interpret the materials they found.

We also proposed this as a roughly cyclical model: we were aware of researchers who look in collections, and then form research questions based on what they know they can find, and so on. It is not necessarily a linear process from question formation, through archive identification and searching to interpretation.

Furthermore we proposed that question formation and identifying archives were the two key processes for users not in higher education. Undergraduate students are given research questions (usually in the form of essay titles or project proposals) and are pointed by their tutors in the direction of the useful library collections. Similarly postgraduate students and academics have (or are developing) skills in identifying sensible, tractable research questions, and know how to the use the library staff and their colleagues to identify likely looking archives. In other words academics and students are a dense group: there are strong and supportive relationships between students, tutors, colleagues and library staff which help them construct research questions and identify useful research materials. Sparsely related non-HE users (we assumed) would have neither the skills or the supportive network. However we assumed that the users would have good skills in searching and interpretation, or at least would have access to tutorials in these skills that our project would not need to replicate.

From this 'why' statement stemmed a set of 'what' statements: that the website should offer a collection of online tutorials on question formation, and a discussion group-like facility to allow interaction between users and library staff to help users identify collections.

A prototype of this system was mocked up and made available to users. We then set out to evaluate the prototype and the assumptions underlying it. This was done by inviting local and family history research groups into the library and visiting meetings of such groups. Individual researchers were also invited into the library to discuss their work with the project team, and public libraries with strong local history sections were contacted and they supplied us with contacts with researchers who used their facilities. We also tried indirect routes to get at possible users: primarily by interviewing archivists about what their collections were used for by non-HE users. Our contact with potential users began to 'span out' from the first users we contacted in exactly the way Wellman predicts the analysis a sparse group would.

Results of evaluating the first iteration. We found that three of the four main assumptions were correct: users did need support identifying archives, and were already competent searching and interpreting archival materials. However we found that, contrary to our expectations, they did have well developed, tractable research questions, or if they did not, then they would not be interested in the collections held in university libraries. In retrospect this makes sense: researchers with badly thought out research questions are likely to be beginners, and would only be interested in the records held in public libraries or in census data. Once the possibilities of the census data and so on have been exhausted, then the researcher may find value in the collections held in university libraries, but by this time they will have become experienced researchers and will have defined and refined their research questions. Note how through this analysis a boundary for our user population has emerged, again, as predicted by Wellman.

Our contact with the archivists also provided a key insight: that what was published about an archival collection described objectively *what was in it*, whereas the archivists told us subjectively what research one *could do* with the collection. This shows the intermediation role that the archivists play (and is discussed in more detail in the next section) but also suggested to us that a better way of supporting users in finding archives useful for their research questions would be encode the archivists' knowledge of what a collection could be used for as a searchable, online database.

Second iteration. Based on the evaluation we then had a much clearer idea of who the users were, what they needed, and how we could fulfill those needs. We now developed a prototype that did not have tutorials on question formation, and had a database of 'use centred descriptions' of University of London archival collections that suggested to personal history researchers what they could do with those collections. This prototype and its underlying assumptions were evaluated, and this time the feedback was much more positive: we now felt we were firmly on track to deliver a useful DL resource.

Third and subsequent iterations. The way that we were to structure these use centred descriptions was determined by further evaluation and iteration, and various user interface issues were dealt with, until a finished artefact was launched at the end of the project.

3.4 Summary

We have shown a design process which is intended not only to design a working artefact, but also to iteratively develop the designers' understanding of the user population. It is not a radical departure from existing methods, but simply makes it explicit that when defining for unbounded groups, the very basic underlying assumptions need to be evaluated and refined as much as the working artefact does.

When looking at the process in action we see that there was a sizeable change in our ideas about the characteristics of the users after the first iteration, and correspondingly the first prototype was completely dropped before entering the second iteration. The key point is that this big change occurred relatively late in the project, but the project managed to cope with it and still deliver a working product on time, largely because we were *expecting* a large change once we had explored enough of the users in our sparse network. This meant that the early decisions and prototypes were held very lightly, and therefore could be abandoned without major cost.

Obviously this description of what actually happened has been retrospectively neatened up. In particular the process of exploring the user networks was not a linear one: to visit local history groups we had to wait until they had meetings, or for personal history researchers to visit the library we had to fit our timetables around their's. This meant that the analysis came in fits and bursts and did not fit neatly into our design iterations.

Another observation that emerged was how important liaison with archivists and other front line library staff was in designing the system. This is because such staff have long been in the job of analysing the needs of their users, and the results of that analysis was very useful to us in designing our DL resources. Even though it may be difficult for system developers working on short term projects to gather requirements directly from users in unbounded networks, it is possible to get a good indirect picture of their needs through librarians and archivists who liaise with them in the long term.

4 Disintermediation in Sparse Networks

Butterworth and Davis Perkins [16] presented an analysis of 'small, specialist libraries' with a focus on how the requirements for developing their digital incarnations differ from those of commercial and academic libraries. In particular they showed that the intermediation roles of the librarian are even more important and extensive for small, specialist libraries. They showed that librarians not only play the intermediation roles between information sources and readers described elsewhere [17, chapter 7], but also play a more social intermediation role between the readers themselves.

In a sparse network the effect of a social intermediator is dramatic: it turns a weakly connected or disconnected network into a much more highly connected network. A sparse network of users may contain several completely separate sub networks, or even completely isolated actors: a social intermediator connects all the sub networks and actors together. In theory, if the intermediator is in contact with all the actors in a network, the effect is to render all the actors at most two relationships away from each other.

This effect is much more profound for a sparse network than for a dense one: for a librarian (or anyone) to play a social intermediation role in a dense network would not dramatically increase the interconnectedness of a dense network, because it is highly interconnected already. There are strong arguments in the literature [18, 19] against disintermediation, and in the case of digital library systems for sparse user networks we contend that disintermediation is particularly detrimental.

In the Accessing our Archival and Manuscript Heritage project it became apparent as we explored our potential users that the main way we could benefit them was helping them to link potential archive with their research questions. This relationship between research question and an archival collection that can be used to address that question is often not clear. Very often an archive can be used in very different ways to the purposes it was collected for. For example London University's School of Oriental and African Studies holds an extensive collection of correspondence sent by 19th Century African missionaries, which has been used by a researcher to create a climate map of Africa in the 19th century. This was possible because the missionaries often wrote home and gave detailed descriptions of the local geography and climate.

ISAD(G) description

Context

Administrative/Biographical history: Patrick Manson was born in 1844 and studied medicine at Aberdeen University, passing M.B. and C.M. in 1865. In 1866 he became medical officer of Formosa for the Chinese imperial maritime customs, moving to Amoy in 1871. Here, while working on elephantoid diseases, he discovered in the tissues of blood-sucking mosquitoes the developmental phase of filaria worms. From 1883 to 1889 he was based in Hong Kong, where he set up a school of medicine that developed into the university and medical school of Hong Kong[. . .]

Content

Scope and content/abstract: Papers of Sir Patrick Manson, 1865-1964, including Manson's diaries, 1865-1879, containing notes on the discovery of mosquitoes as carriers of malaria and patient case notes; bound manuscript notes of his discovery of filaria, 1877; original drawings of eggs of bilharzias and embryos of guinea worms, 1893; drawings by Manson of filarial embryos, 1891; correspondence with Charles Wilberforce Daniels[. . .]

Use centred description

Detailed usage description

The London School of Hygiene and Tropical Medicine holds an archive of the medical examinations of people who emigrated to the British colonies and protectorates between 1898 and 1919. As well as giving a detailed account of the subject's health, each record gives a small amount of family history parents, children and siblings) as well as some details about their current job, the job that they were intending to take up in the colonies and its location.

If you have a relative who apparently 'disappeared' at the end of the 19th Century, e.g. they're in the 1891 census, but not in the 1901 census, they may have emigrated, and this collection may give you a clue as to where and when they went[. . .]

How to tell if the collection is useful

If you know that a family member emigrated between 1898 and 1919 then this collection is clearly useful. If you don't know for sure, but suspect that you may have an ancestor who emigrated, you may email a query to LSHTM's archivist, giving as much detail as possible[. . .]

Fig. 1. A partial ISAD(G) [20] collection level description and partial use centred description of the Sir Patrick Manson archive held at the London School of Hygiene and Tropical Medicine (Both descriptions are edited for size)

A potential problem for a researcher is that the published archival descriptions objectively describe what is in an archive, who created it and when, but do not describe what can be done with the collection. To identify what a collection can be used for takes either lateral thinking, a lucky guess, or intermediation by the archivists who know what uses their collections have been put to in the past and can pass this knowledge on to other researchers. This social intermediation role of passing knowledge from researcher to researcher is crucial in sparse networks; if one researcher works out that archive X can be used for purpose Y, this knowledge is not likely to propogate around a sparse network without intermediation.

The disparity between what a published description of an archival collection says, (ie. what is in the collection), and what an archivist will tell you about their collections, (ie. what you can do with a collection) became one of the central points of the project. We set about interviewing archivists about what personal history researchers can use their collections for, and published these as a set of 'use centred descriptions' on the site developed by the project. (See the 'Helpers' site: `http://helpers.shl.lon.ac.uk/`.)

Again space precludes a full review of use centred descriptions (see [15, Section 6]) but as an example figure 1 shows part of a use centred description and the standard archival description of the same collection. The collection described is the Sir Patrick Manson archive held at the London School of Hygiene and Tropical Medicine. Sir Patrick was a founder of the School and was instrumental in showing that malaria was transmitted by mosquitoes. His working papers form the basis of this archival collection. The standard archival description details Sir Patrick's life and lists the different materials held in his archive. If you are a family history researcher there is no indication in the description that the collection would be of any use to you. However the use centred description shows that in the collection is a list of medical examinations of twelve thousand people who emigrated to the British colonies between 1898 and 1919. As well as containing medical data these records also detailed where the subjects were going in the colonies to work and their immediate family. This information could be vital to family historians.

The common approach to repairing disintermediation gaps in digital libraries is to allow users direct contact with library staff, via email, discussion groups and chat rooms. (For example, see the People's network Enquire service[3].) We attempted the same end by encoding the archivists' knowledge about which archives are useful to which users, and making it available online as a searchable database. Clearly no-one would claim that this is a replacement for the job that archivists and other front end library staff do, but it is a way of allowing knowledge about the uses that an archive can be put to to travel through a sparse network of researchers.

5 Conclusions

This paper has argued that there are classes of potential digital library users outside of tertiary education who are best characterised as sparse, unbounded networks. We argue that most requirements engineering techniques make the assumption that a system is designed for a bounded group of users, and therefore do not serve DL development well. Furthermore we have contended that intermediation is particularly important in a sparse network, and we have discussed how 'use centred descriptions' of archival collections act as an intermediation tool.

We believe that sparse, unbounded networks are much more common within traditional and digital library users than is suggested by the concentration on dense, bounded networks reported in the DL literature. We would propose further work, particularly looking at users of public libraries, to further draw out the characteristics of these user networks.

References

1. Shen, R., Gonçalves, M.A., Fan, W., Fox, E.: Requirements gathering and modelling of domain-sprecific digital libraries with the 5S framework: an archaeological case study with ETANA. [21] 1–12
2. Monroy, C., Furuta, R., Mallen, E.: Visualising and exploring picasso's world. In Marshall, C., Henry, G., Delcambre, L., eds.: Proceedings of 2003 Joint Conference on Digital Libraries, IEEE (2003)

[3] http://www.peoplesnetwork.gov.uk/enquire/

3. Wilson, R., Landoni, M., Gibb, F.: Guidelines for designing electronic books. In Agosti, M., Thanos, C., eds.: Proceedings of 6th European Conference on Digital Libraries, ECDL 2002. Volume LNCS 2458., Springer Verlag (2002)
4. Wellman, B.: An electronic group is virtually a social network. In Kiesler, S., ed.: Culture of the internet. Lawrence Erlbaum (1997) 179–205
5. Wasserman, S., Faust, K.: Social network analysis: methods and applications. Cambridge University Press (1994)
6. Wellman, B., Berkowitz, S.D., eds.: Social structures: a network approach. JAI Press Inc. (1988)
7. Davis Perkins, V., Butterworth, R., Curzon, P., Fields, B.: A study into the effect of digitisation projects on the management and stability of historic photograph collections. [21] 278–289
8. Keil, M., Cule, P., Lyytinen, K., Schmidt, R.: A framework for identifying software project risks. Commun. ACM **41**(11) (1998) 76–83
9. McGraw, K., Harbison, K.: User-centred requirements: the scenario based engineering process. Lawrence Erlbaum Associates (1997)
10. Sommerville, I., Sawyer, P.: Requirements engineering: a good practice guide. Wiley (1997)
11. Hix, D., Hartson, H.R.: Developing user interfaces: Ensuring usability through product and process. John Wiley and Sons (1993)
12. Dix, A., Finlay, J., Abowd, G., Beale, R.: Human computer interaction. Third edn. Pearson (2004)
13. Boehm, B.: A spiral model of software development and enhancement. IEEE Computer **21**(5) (1988) 61–72
14. Yu, E.: Towards modelling and reasoning support for early-phase requirements engineering. In: Proceedings of RE-97: 3rd International Symposium on Requirements Engineering. (1997) 226–235
15. Butterworth, R.: The Accessing our Archival and Manuscript Heritage project and the development of the ‘Helpers’ website. Technical Report IDC-TR-2006-001, Interaction Design Centre, School of Computing Science, Middlesex University (2006) Available from `http://www.cs.mdx.ac.uk/research/idc/tech_reports.html`.
16. Butterworth, R., Davis Perkins, V.: Assessing the roles that a small specialist library plays to guide the development of a hybrid digital library. In Crestani, F., Ruthven, I., eds.: Information Context: Nature, Impact, and Role: 5th International Conference on Conceptions of Library and Information Sciences, CoLIS 2005. Volume LNCS 3507., Springer Verlag (2005) 200–211
17. Nardi, B.A., O’Day, V.: Information Ecologies. MIT Press (2000)
18. Borgman, C.: From Gutenberg to the global information infrastructure: access to information in the networked world. MIT Press (2000)
19. Vishik, C., Whinston, A.: Knowledge sharing, quality, and intermediation. In: WACC ’99: Proceedings of the international joint conference on Work activities coordination and collaboration, ACM Press (1999) 157–166
20. International Council on Archives: ISAD(G): General international standard archival description. (1999) Second edition.
21. Rauber, A., Christodoulakis, S., Min Tjoa, A., eds.: Research and Advanced Technology for Digital Libraries: 9th European Conference, ECDL 2005. In Rauber, A., Christodoulakis, S., Min Tjoa, A., eds.: Research and Advanced Technology for Digital Libraries: 9th European Conference, ECDL 2005. (2005)

Design and Selection Criteria for a National Web Archive

Daniel Gomes, Sérgio Freitas, and Mário J. Silva

University of Lisbon, Faculty of Sciences
1749-016 Lisboa, Portugal
dcg@di.fc.ul.pt, sfreitas@lasige.di.fc.ul.pt, mjs@.di.fc.ul.pt

Abstract. Web archives and Digital Libraries are conceptually similar, as they both store and provide access to digital contents. The process of loading documents into a Digital Library usually requires a strong intervention from human experts. However, large collections of documents gathered from the web must be loaded without human intervention. This paper analyzes strategies to select contents for a national web archive and proposes a system architecture to support it.[1]

1 Introduction

Publishing tools, such as Blogger, enabled people with limited technical skills to become web publishers. Never before in the history of mankind so much information was published. However, it was never so ephemeral. Web documents such as news, blogs or discussion forums are valuable descriptions of our times, but most of them will not last longer than one year [21] If we do not archive the current web contents, the future generations could witness an information gap in our days. The archival of web data is of interest beyond historical purposes. Web archives are valuable resources for research in Sociology or Natural Language Processing. Web archives could also provide evidence in judicial matters when ephemeral offensive contents are no longer available online. The archival of conventional publications has been directly managed by human experts, but this approach can not be directly adopted to the web, given its size and dynamics. We believe that web archiving must be performed with minimum human intervention. However, this is a technologically complex task. The Internet Archive collects and stores contents from the world-wide web. However, it is difficult for a single organization to archive the web exhaustively while satisfying all needs, because the web is permanently changing and many contents disappear before they can be archived. As a result, several countries are creating their own national archives to ensure the preservation of contents of historical relevance to their cultures [6].

Web archivists define boundaries of national webs as selection criteria. However, these criteria influence the coverage of their archives. In this paper, we

[1] This study was partially supported by FCT under grant SFRH/BD/11062/2002 (scholarship) and FCCN.

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 196–207, 2006.

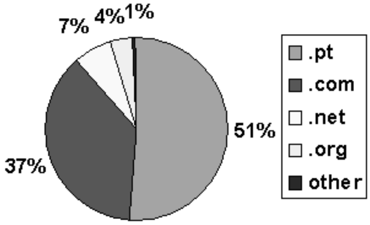


Fig. 1. Distribution of documents per domain from the Portuguese web

analyze strategies for selecting contents for a national web archive and present a system's architecture to support a web archiving system. This architecture was validated through a prototype named Tomba. We loaded Tomba with 57 million documents (1.5 TB) gathered from the Portuguese web during 4 years to update the indexes of a search engine and made this information publicly available through a web interface (available at tomba.tumba.pt). The main contributions of this paper are: i) the evaluation of selection strategies to populate a web archive; ii) a system's architecture to support a web archive.

In the following Section we discuss strategies to populate a web archive. In Section 3, we present the architecture of the Tomba prototype. Section 4 presents related work and in Section 5 we conclude our study and propose future work.

2 Selecting

Web archivists define strategies to populate web archives according to the scope of their actions and the resources available. An archive can be populated with contents delivered from publishers or harvested from the web. The delivery of contents published on the web works on a voluntary basis in The Netherlands but it is a legislative requirement in Sweden [20]. However, the voluntary delivery of contents is not motivating for most publishers, because it requires additional costs without providing any immediate income. On the other hand, it is difficult to legally impose the delivery of contents published on sites hosted on foreign web servers, outside a country's jurisdiction. The absence of standard methods and file formats to support the delivery of contents is also a major drawback, because it inhibits the inclusion of delivery mechanisms in popular publishing tools. Alternatively, a web archive can be populated with contents periodically harvested from the country's web. However, defining the boundaries of a national web is not straightforward and the selection policies are controversial.

We used the Portuguese web as a case study of a national web and assumed that it was composed by the documents hosted on a site under the .PT domain or written in the Portuguese language hosted in other domains, linked from .PT [10]. We used a crawl of 10 million documents harvested from the Portuguese web in July, 2005 as baseline to compare the coverage of various selection policies.

2.1 Country Code Top Level Domains

There are two main classes of top-level domains (TLD): generic (gTLDs) and country code (ccTLDs). The gTLDs were meant to be used by particular classes of organizations (e.g. COM for commercial organizations) and are administrated by several institutions world wide. The ccTLDs are delegated to designated managers, who operate them according to local policies adapted to best meet the economic, cultural, linguistic, and legal circumstances of the country. Hence, sites with a domain name under a ccTLD are strong candidates for inclusion in a web archive. However, this approach excludes the documents related to a country hosted outside the ccTLD. Figure 1 presents the distribution of documents from the Portuguese web per domain and shows that 49% of its documents are hosted outside the ccTLD .PT.

2.2 Exclude Blogs

Blogs have been introduced as frequent, chronological publications of personal thoughts on the web. Although the presence of blogs is increasing, most of them are rarely seen and quickly abandoned. According to a survey, "the typical blog is written by a teenage girl who uses it twice a month to update her friends and classmates on happenings on her life" [5], which hardly matches the common requirements of a document with historical relevance. On the other hand, blogs are also used to easily publish and debate any subject, gaining popularity against traditional web sites. Blogs that describe the life of citizens from different ages, classes and cultures will be an extremely valuable resource for a description of our times [8].

We considered that a site is a blog if it contained the string "blog" on the site name and observed that 15.5% of the documents in the baseline would have been excluded from a national web archive if blogs were not archived. 67% of the blog documents were hosted under the .com domain and 33% were hosted on blogs under the .PT domain. One reason we found for this observation is that most popular blogging sites are hosted under the .COM domain, which tends to increase the number of documents from a national web hosted outside the country code TLD (Blogspot that holds 63% of the Portuguese blogs).

2.3 Physical Location of Web Servers

The RIPE Network Management Database provides the country where an IP address was firstly allocated or assigned. One could assume that the country's web is composed by the documents hosted on servers physically located on the country. We observed that only 39.4% of the IP addresses of the baseline Portuguese web were assigned to Portugal.

2.4 Select Media Types

A web archive may select the types of the contents it will store depending on the resources available and the scope of the archive. For instance, one may populate

Table 1. Prevalence of media types on the Portuguese web

MIME type	**avg size (KB)**	**%docs.**
text/html	24	**61.2%**
image/jpeg	32	**22.6%**
image/gif	9	**11.4%**
application/pdf	327	1.6%
text/plain	102	0.7%
app'n/x-shockwave-flash	98	0.4%
app'n/x-tar	**1,687**	0.1%
audio/mpeg	**1,340**	0.04%
app'n/x-zip-compressed	**541**	0.1%
app'n/octet-stream	454	0.1%
other	129	1.8%

a web archive exclusively with audio contents. Preservation strategies must be implemented according to the format of the documents. For instance, preserving documents in proprietary formats may require having to preserve also the tools to interpret them. The costs and complexity of the preservation of documents increases with the variety of media types archived and it may become unbearable. Hence, web archivists focus their efforts on the preservation of documents with a selected set of media types. Table 1 presents the coverage of selection strategies according to the selected media types. We can observe that a web archive populated only with HTML pages, JPEG and GIF images covers 95.2% of a national web.

2.5 Ignore Robots Exclusion Mechanisms

The Robots Exclusion Protocol (REP) enables authors to define which parts of a site should not be automatically harvested by a crawler through a file named "robots.txt" [16] and the meta-tag ROBOTS indicates if a page can be indexed and the links followed [26].Search engines present direct links to the pages containing relevant information to answer a given query. Some publishers only allow the crawl of the site's home page to force readers to navigate through several pages containing advertisements until they find the desired page, instead of finding it directly from search engine results. One may argue that archive crawlers should ignore these exclusion mechanisms to achieve the maximum coverage of the web. However, the exclusion mechanisms are also used to prevent the crawling of sites under construction and infinite contents such as online calendars [24]. Moreover, some authors create spider traps, that are sets of URLs that cause the infinite crawl of a site [15], to punish the crawlers that do not respect the exclusion mechanisms. So, ignoring the exclusion mechanisms may degrade the performance of an archive crawler.

We observed that 19.8% of the Portuguese web sites contained the "robots.txt" file but the REP forbade the crawl of just 0.3% of the URLs. 10.5% of the pages contained the ROBOTS meta-tag but only 4.3% of them forbade the indexing of

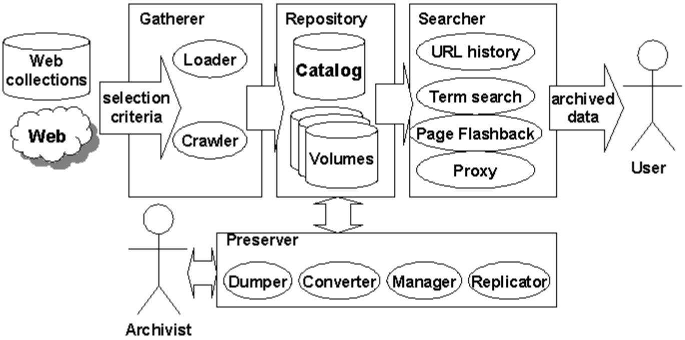


Fig. 2. Architecture of the Tomba web archive

the page and 5% disallowed the following of links. The obtained results suggest that ignoring exclusion mechanisms does not significantly increase the coverage of a national web crawl. However, this behavior may degrade the crawler's performance because exclusion mechanisms are also used to prevent crawlers against hazardous situations.

3 The Tomba Web Archive

The Tomba web archive is a prototype system developed at the University of Lisbon to research web archiving issues. A web archive system must present an architecture able to follow the pace of the evolution of the web, supporting distinct selection criteria and gathering methods. Meta-data must be kept to ensure the correct interpretation and preservation of the archived data. A collection of documents built through incremental crawls of the web contains duplicates, given the documents that remain unchanged and the different URLs that reference the same document. It is desirable to minimize duplication among the archived data to save storage space without jeopardizing performance. The storage space must be extensible to support the collection growth and support various storage policies according to the formats of the archived documents and the level of redundancy required. The archived data should be accessible to humans and machines, supporting complementary access methods to fulfill the requirements of distinct usage contexts. There must be adequate tools to manage and preserve the archived documents, supporting their easy migration to different technological platforms.

Figure 2 represents the architecture of Tomba. There are 4 main components. The *Gatherer* is responsible for collecting web documents and integrating them in the archive. The *Repository* stores the contents and their correspondent meta-data. The *Preserver* provides tools to manage and preserve the archived data. The *Searcher* allows human users to easily access the archived data. The *Archivist* is a human expert that manages preservation tasks and defines selection criteria to automatically populate the archive.

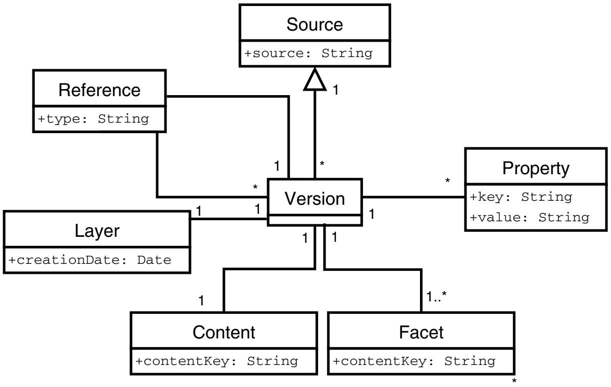


Fig. 3. Data model

3.1 Repository

A content is the result of a successful download from the web (e.g. an HTML file), while meta-data is information that describes it (e.g. size). The *Repository* is composed by the *Catalog* [3] that provides high performance structured access to meta-data and the *Volumes* [9] that provide an extensible storage space to keep the contents, eliminating duplicates among them.

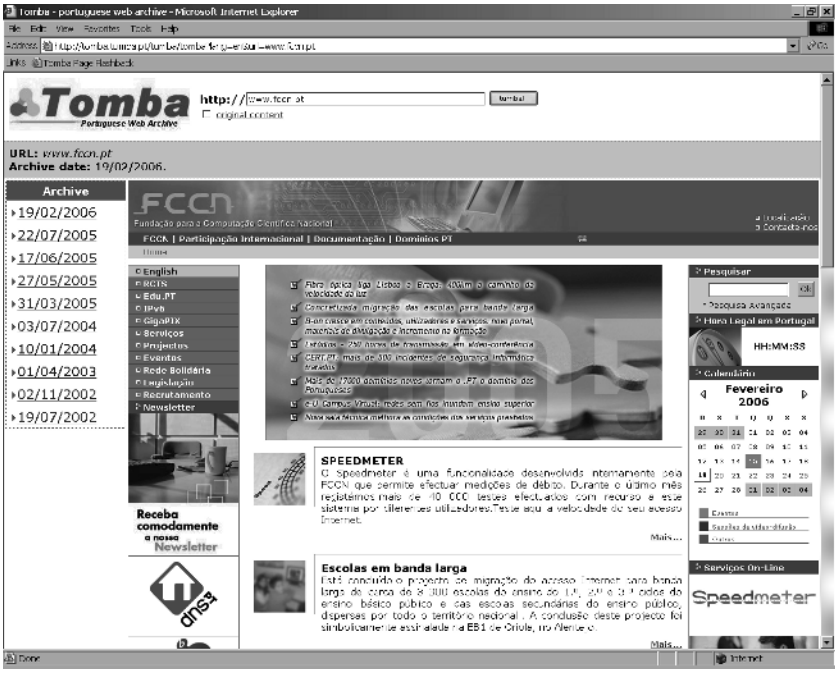
Figure 3 describes the data model of the Catalog. We assume that the archive is loaded in bulk with snapshots of the web. The *Source* class identifies the origin of the document, for example an URL on the web. Each *Version* represents a snapshot of the information gathered from a Source. The Versions correspondent to the same snapshot of the web are aggregated in *Layers.* A Layer represents the time interval from its creation until the creation of the next one. This way, time is represented in a discrete fashion within the archive, facilitating the identification of web documents that need to be presented together, such as a page and the embedded images. The Property class holds property lists containing meta-data related to a Version. The use of property lists instead of a static meta-data model, enables the incremental annotation of contents with meta-data items when required in the future. The *Content* and *Facet* classes reference documents stored in the Volumes. The former references the documents in their original format and the latter alternative representations. For instance, a Content is an HTML page that has a Facet that provides the text contained in it. In the archive, Facets provide storage for current representations of contents retrieved earlier in obsolete formats. The Repository supports merging the Content, Facets and meta-data of a Version into a single Facet in a semi-structured format (XML), so that each document archived in a Volume can be independently accessed from the Catalog. There are web documents that contain useful information to preserve other ones. For instance, a web page containing the specification of the HTML format could be used in the future to interpret documents written in this format. The *Reference* class enables the storage of associations of Versions that are related to each other.

3.2 Gatherer

The *Gatherer*, composed by the *Loader* and the *Crawler*, integrates web data in the Repository. The *Loader* was designed to support the delivery of web contents by publishers and receive previously compiled collections of documents. The *Crawler* iteratively harvests information from the web, downloading pages and following the linked URLs. Ideally, a page and the embedded or referenced documents would be crawled sequentially to avoid that some of them become unavailable meanwhile. Sequentially crawling all the documents referenced by a page degrades the crawler's performance, because harvesting the documents hosted outside a site requires additional DNS lookups and establishment of new TCP connections. According to Habib and Abrams, these two factors account for 55% of the time spent downloading web pages [12]. Crawling the documents of one site at a time in a breadth first mode and postponing the crawl of external documents until the corresponding sites are visited, is a compromise solution that ensures that the majority (71%) of the embedded documents internal to each site are crawled in a short notice, without requiring additional bandwidth usage [18].

3.3 Preserver

Replication is crucial to prevent data loss and ensure the preservation of the archived documents. The replication of data among mirrored storage nodes must consider the resources available, such as disk throughput and network bandwidth. A new document loaded into the archive can be immediately stored across several mirrors, but this is less efficient than replicating documents in bulk. Considering that an archive is populated with documents crawled from the web within a limited time interval, the overhead of replicating each document individually could be prohibitive. The *Replicator* copies the information kept in a Volume to a mirror in batch after each crawl is finished. The *Dumper* exports the archived data to a file using 3 alternative formats: i) WARC, proposed by the Internet Archive to facilitate the exportation of data to other web archives [17]; ii) an XML based format to enable flexible automatic processing; iii) a textual format with minimum formatting created to minimize the space used by the dump file. The dissemination of the archived documents as public collections is an indirect way to replicate them outside the archive, increasing their chance of persisting into the future. These collections are interesting for scientific evaluations [14] or to be integrated in other web archives. The main obstacles to the distribution of web collections are their large size, the lack of standards to format them in order to be easily integrated in external systems and copyright legislation that requires authorization from the authors of the documents to distribute them. Obtaining these authorizations is problematic for web collections having millions of documents written by different authors. The archived documents in obsolete formats must be converted to up-to-date formats to maintain their contents accessible. The *Converter* iterates through the documents kept in the Repository and generates Facets containing alternative representations in different formats. The *Manager* allows a human user



URL: *www.fccn.pt*
Archive date: 19/02/2006.

- 19/02/2006
- 22/07/2005
- 17/06/2005
- 27/05/2005
- 31/03/2005
- 03/07/2004
- 10/01/2004
- 01/04/2003
- 02/11/2002
- 19/07/2002

SPEEDMETER

O Speedmeter é uma funcionalidade desenvolvida internamente pela FCCN que permite efectuar medições de débito. Durante o último mês registámos mais de 40 000 testes efectuados com recurso a este sistema por diferentes utilizadores.Teste aqui a velocidade do seu acesso Internet.

Escolas em banda larga

Está concluído o projecto de migração do acesso Internet para banda larga de cerca de 8 300 escolas do ensino do 1.º, 2.º e 3.º ciclos do ensino básico público e das escolas secundárias do ensino público, dispersas por todo o território nacional . A conclusão deste projecto foi simbolicamente assinalada na EB1 de Criole, no Alentejo.

Fig. 4. Tomba web interface

to access and alter the archived information. The meta-data contained in the *Content-Type* HTTP header field identifies the media type of a web document but sometimes it does not correspond to the real media type of the document. On our baseline crawl, 1.8% of the documents identified as plain text were in fact JPEG image files. The format of a document is commonly related to the file name extension of the URL that references it. This information can be used to automatically correct erroneous media type meta-data. However, the usage of file name extensions is not mandatory within URLs and the same file name extension may be used to identify more than 1 format. For example, the extension .rtf identifies documents in the application/rtf and text/richtext media types. In these cases, a human expert can try to identify the media type of the document and correct the corresponding meta-data using the Manager.

3.4 Searcher

The *Searcher* provides 3 methods for accessing the archived data: *Term Search*, *URL History* or *Navigation*. The Term Search method finds documents containing a given term. The documents are previously indexed to speed up the searches. The URL History method finds the versions of a document referenced by an URL. The Navigation method enables browsing the archive using a web proxy.

Figure 4 presents the public web interface of Tomba that supports the URL History access method. Navigation within the archive begins with the submission of an URL in the input form of the Tomba home page. In general, multiple different URLs reference the same resource on the web and it may seem indifferent to users to submit any of them. If only exact matches on the submitted URL were accepted, some documents might not be found in the archive. Hence, Tomba expands each submitted URL to a set of URLs that are likely to reference the same resource, and then searches for them. For instance, if a user inputs the URL www.tumba.pt, Tomba will look for documents harvested from the URLs: www.tumba.pt/, tumba.pt, www.tumba.pt/index.html, www.tumba.pt/index.htm, www.tumba.pt/index.php, www.tumba.pt/index.asp. On the visualization interface, the archive dates of the available versions of a document are displayed on the left frame. The most recent version of the document is initially presented on the right frame and users can switch to other versions by clicking on the associated dates. The versions presented on the left frame enable a quick tracking of the evolution of a document. The documents harvested from the web are archived in their original format. However, they are transformed before being presented to the user to enable mimicking their original layout and allow a user to follow links to other documents within the archive when activating a link on a displayed page. The documents are parsed and the URLs to embedded images and links to other documents are replaced to reference archived documents. When a user clicks on a link, Tomba picks the version of the URL in the same layer of the referrer document and displays it on the right frame along with the correspondent versions on the left frame. A user may retrieve an archived document without modifications by checking the box *original content* below the submission form (Figure 4). This is an interesting feature for authors that want to recoverer old versions of a document. The Page Flashback mechanism enables direct access to the archived versions of a document from the web being displayed on the browser. The user just needs to click on a toolbar icon and the versions of the page archived in Tomba will be immediately presented.

The URL History access method has 3 main limitations. First, users may not know which URL they should submit to find the desired information. Second, the short life of URLs limits their history to a small number of versions. The Tomba prototype was loaded with 10 incremental crawls of the Portuguese web but on average each URL referenced just 1.7 versions of a document. Third, the replacement of URLs may not be possible in pages containing format errors or complex scripts to generate links. If these URLs reference documents that are still online, the archived information may be presented along with current documents. The Term Search and Navigation complement the URL History but they have other limitations. The Term Search finds documents independently from URLs but some documents may not be found because the correspondent text could not be correctly extracted and indexed [7] The Navigation method enables browsing the archive without requiring the replacement of URLs because all the HTTP requests issued by the user's browser must pass through the proxy that returns contents only for archived documents. However, it might be hard to find the desired information by following links among millions of documents.

4 Related Work

According to the National Library of Australia there are 16 countries with well-established national Web archiving programs [20]. The Internet Archive was the pioneer web archive. It has been executing broad crawls of the web and released an open-source crawler named Heritrix [11]. The National Library of Australia founded its web archive initiative in 1996 [22].It developed the PANDAS (PANDORA Digital Archiving System) software to periodically archive Australian online publications, selected by librarians for their historical value. The British Library leads a consortium that is investigating the issues of web archival [4]. The project aims to collect and archive 6,000 selected sites from the United Kingdom during 2 years using the PANDAS software. The sites have been stored, catalogued and checked for completeness. The MINERVA (Mapping the INternet Electronic Resources Virtual Archive) Web Archiving Project was created by the Library of the Congress of the USA and archives specific publications available on the web that are related to important events, such as an election [25].

In December 2004 the Danish parliament passed a new legal deposit law that calls for the harvesting of the Danish part of the Internet for the purpose of preserving cultural heritage and two libraries became responsible for the development of the Netarkivet web archive [19].The legal deposit of web contents in France will be divided among the Institut National de l'Audiovisuel (INA) and the National Library of France (BnF). Thomas Drugeon presented a detailed description of the system developed to crawl and archive specific sites related to media and audiovisual [7]. The BnF will be responsible for the archive of online writings and newspapers and preliminary work in cooperation with a national research institute (INRIA) has already begun [1].

The National Library of Norway had a three-year project called Paradigma (2001-2004) to find the technology, methods and organization for the collection and preservation of electronic documents, and to give the National Library's users access to these documents [2]The defunct NEDLIB project (1998-2000) included national libraries from several countries (including Portugal) and had the purpose of developing harvesting software specifically for the collection of web resources for an European deposit library [13].The Austrian National Library together with the Department of Software Technology at the Technical University of Vienna, initiated the AOLA project (Austrian On-Line Archive) [23].The goal of this project is to build an archive by harvesting periodically the Austrian web. The national libraries of Finland, Iceland, Denmark, Norway and Sweden participate in the Nordic Web Archive (NWA) project [?] The purpose of this project is to develop an open-source software tool set that enables the archive and access to web collections.

5 Conclusions and Future work

We proposed and evaluated selection criteria to automatically populate a national web archive. We observed that no criteria alone provides the solution for selecting the contents to archive and combinations must be used. Some criteria are not selective but their use may prevent difficulties found while populating the archive. In particular, we conclude that populating a national web archive only with documents hosted in sites under the country's Top Level Domain or physically located on the country excludes a large amount of documents. The costs and complexity of the preservation of documents increases with the variety of media types archived. We observed that archiving documents of just three media types (HTML, GIF and JPEG) reduced the coverage of a national web only 5%. We conclude that this is an interesting selection criterion to simplify web archival, in exchange for a small reduction on the coverage of the web.

We described the architecture of an information system designed to fulfil the requirements of web archiving and validate it through the development of a prototype named Tomba. We loaded Tomba with 57 million documents (1.5 TB) harvested from the Portuguese web during the past 4 years and explored three different access methods. None of them is complete by itself, so they must be used in conjunction to provide access to the archived data.

As future work, we intend to enhance accessibility to the archived information by studying an user interface suitable to access a web archive.

References

1. S. Abiteboul, G. Cobena, J. Masanes, and G. Sedrati. A first experience in archiving the french web. In *ECDL '02: Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*, pages 1–15, London, UK, 2002. Springer-Verlag.
2. K. Albertsen. The paradigma web harvesting environment. In *Proceedings of 3rd ECDL Workshop on Web Archives*, Trondheim, Norway, August 2003.
3. J. Campos. Versus: a web repository. Master thesis, 2003.
4. U. W. A. Consortium. Uk web archiving consortium: Project overview. http://info.webarchive.org.uk/, January 2006.
5. P. D. Corporation. Perseus blog survey. September 2004.
6. M. Day. Collecting and preserving the world wide web. `http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf`, 2003.
7. T. Drugeon. A technical approach for the french web legal deposit. In *5th International Web Archiving Workshop (IWAW05)*, Viena, Austria, September 2005.
8. R. Entlich. Blog today, gone tomorrow? preservation of weblogs. *RLG Diginews*, 8(4), August 2004.
9. D. Gomes, A. L. Santos, and M. J. Silva. Managing duplicates in a web archive. In L. M. Liebrock, editor, *Proceedings of the 21th Annual ACM Symposium on Applied Computing (ACM-SAC-06)*, Dijon, France, April 2006.
10. D. Gomes and M. J. Silva. Characterizing a national community web. *ACM Trans. Inter. Tech.*, 5(3):508–531, 2005.
11. M. S. I. R. Gordon Mohr, Michele Kimpton. Introduction to heritrix, an archival quality web crawler. In *4th International Web Archiving Workshop (IWAW04)*, Bath, UK, September 2004. Internet Archive, USA.
12. M. A. Habib and M. Abrams. Analysis of sources of latency in downloading web pages. In *WebNet*, San Antonio, Texas, USA, November 2000.
13. J. Hakala. Collecting and preserving the web: Developing and testing the nedlib harvester. *RLG Diginews*, 5(2), April 2001.

14. D. Hawking and N. Craswell. Very large scale retrieval and web search. In E. Voorhees and D. Harman, editors, *The TREC Book*. MIT Press, 2004.
15. A. Heydon and M. Najork. Mercator: A scalable, extensible web crawler. *World Wide Web*, 2(4):219–229, 1999.
16. M. Koster. A standard for robot exclusion. http://www.robotstxt.org/wc/norobots.html, June 1994.
17. J. Kunze, A. Arvidson, G. Mohr, and M. Stack. *The WARC File Format (Version 0.8 rev B)*, January 2006.
18. M. Marshak and H. Levy. Evaluating web user perceived latency using server side measurements. *Computer Communications*, 26(8):872–887, 2003.
19. F. McCown. Dynamic web file format transformations with grace. In *5th International Web Archiving Workshop (IWAW05)*, Viena, Austria, September 2005.
20. National Library of Australia. Padi - web archiving. http://www.nla.gov.au/padi/topics/92.html, January 2006. 18.
21. A. Ntoulas, J. Cho, and C. Olston. What's new on the web?: the evolution of the web from a search engine perspective. In *Proceedings of the 13th international conference on World Wide Web*, pages 1–12. ACM Press, 2004.
22. M. Phillips. PANDORA, Australia's Web Archive, and the Digital Archiving System that Supports it. *DigiCULT.info*, page 24, 2003.
23. A. Rauber, A. Aschenbrenner, and O. Witvoet. Austrian on-line archive processing: Analyzing archives of the world wide web, 2002.
24. H. Snyder and H. Rosenbaum. How public is the web?: Robots, access, and scholarly communication. Working paper WP-98-05, Center for Social Informatics, Indiana University, Bloomington, IN USA 47405-1801, January 1998.
25. The Library of Congress. Minerva home page (mapping the internet electronic resources virtual archive, library of congress web archiving). http://lcweb2.loc.gov/cocoon/minerva/html/minerva-home.html, January 2006.
26. The Web Robots Pages. Html author's guide to the robots meta tag. http://www.robotstxt.org/wc/meta-user.html, March 2005.

What Is a Successful Digital Library?

Rao Shen, Naga Srinivas Vemuri, Weiguo Fan, and Edward A. Fox

Digital Library Research Laboratory, Virginia Tech, USA
{rshen, nvemuri, wfan, fox}@vt.edu

Abstract. We synthesize diverse research in the area of digital library (DL) quality models, information systems (IS) success and adoption models, and information-seeking behavior models, to present a more integrated view of the concept of DL success. Such a multi-theoretical perspective, considering user community participation throughout the DL development cycle, supports understanding of the social aspects of DLs and the changing needs of users interacting with DLs. It also helps in determining when and how quality issues can be measured and how potential problems with quality can be prevented.

1 Introduction

Hundreds of millions of dollars have been invested since the early 1990s in research and development related to digital libraries (DLs). Further R&D is needed worldwide [17] if the tremendous potential of DLs is to be achieved. Hence, determining the key characteristics of DL success is of the utmost importance.

What qualifies as a successful DL, and what does not? As this question begins to be analyzed, more questions arise. Who is the intended user of a DL? What is the user's goal for using the DL? What are individual organizations trying to get from their DLs?

For several years, researchers from various disciplines have studied different perspectives of DL success and have generated many interesting yet often isolated findings. Some findings have provided different although sometime overlapping perspectives on how to evaluate DLs. One of them is the DL quality model developed by Gonçalves [11]. For each key concept of a minimal DL, [11] lists a number of dimensions of quality and a set of numerical measurements for those quality dimensions.

Though many would consider a DL to be a type of information system (IS), it often is forgotten that there is a long tradition in IS research of evaluating the success of a generic IS. A variety of measures have been used. Two primary research streams, the user satisfaction literature and the technology acceptance literature (i.e., the technology acceptance model, or TAM) have been investigated. User satisfaction is based on users' attitudes toward a system. We define satisfaction as a user's affective state presenting an emotional reaction to an entire DL and the consequence of the user's experiences during various information-seeking stages. Therefore, we seek to understand the changing needs of users interacting with the DL, and the users' information-seeking behavior during these stages [1]. Fortunately, too, information-seeking behavior has been studied for decades, and many models have been generated.

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 208 – 219, 2006.

A system succeeds when its intended users use it as frequently as needed. User satisfaction prompts user acceptance of the system and leads to higher system usage, because attitude leads to action. Thus, DL user satisfaction can lead to DL success.

The rest of this paper is organized as follows. Section 2 presents the background for our proposed model, which is described in Section 3. Section 4 presents a case study of our model in a domain specific DL. Section 5 concludes the paper.

2 Prior Work

Library and information science researchers, such as those attending the workshop on "Evaluation of Digital Libraries," have investigated the evaluation of DLs [2, 18]. Saracevic [21] was one of the first to consider the problem. According to his analysis, there are no clear agreements regarding the elements of criteria, measures, and methodologies for DL evaluation. The challenge is made more complex by the various classes of users [4]. In an attempt to fill some gaps in this area, Fuhr et al. [10] proposed a description scheme for DLs based on four dimensions. However, a focus on usability of DLs has lagged, especially regarding the non-user-oriented technical topics in the DL literature. There are a few reported studies: inspection of NCSTRL was described in [13]; evaluation of the ACM, IEEE-CS, NCSTRL, and NDLTD digital libraries was reported in [15]; evaluations of ADL and ADEPT were documented in [14] and [6], respectively.

Theories regarding DLs, IS success and adoption, and information-seeking behavior have evolved in parallel. They provide foundations that can be integrated to help answer the question: what is a successful DL? The prior research suggests the need for a more comprehensive view of DL success. There also have been calls for research to empirically validate and extend IS success and adoptions models into varying contexts [25]. Motivated by these calls for research and the increasing number of DL users with varying skills and from different backgrounds and cultures, we seek to answer the question: what is the appropriate model of DL success from the perspective of end users (DL patrons)?

DLs are complex information systems; therefore, research on generic IS may be applied to DLs. The most prominent IS success models existing in the literature today are by Venkatesh [25], DeLone [7], and Seddon [22]. They are discussed in subsections 2 and 3 below. But first we should consider how system usage relates to success.

1. System Usage as a Success Measure

System usage has been considered to be an important indicator of IS success in a number of empirical studies, for many systems. However, simply measuring the amount of time a system is used does not fully capture the relationship between usage and the realization of expected results. The nature, extent, quality, and appropriateness of the system use also should be considered. The nature of system use should be addressed by determining whether the full functionality of a system is being used for the intended purpose. Accordingly, we believe that log analysis could be beneficial to the measurement of DL usage.

2. Technology Acceptance Model (TAM): Predict Intention to Use

TAM provides predictions of intention to use by linking behaviors to attitudes that are consistent with system usage, in time, target, and context. Venkatesh's model [25] predicted behavioral intention to use a system and is a unified model of the eight most popular behavioral IT acceptance theories in the literature. It consists of four core determinants of intention and usage, as shown in Fig. 1. They are: performance expectancy, effort expectancy, social influence, and facilitating conditions.

Despite its predictive ability, TAM provides only limited guidance about how to influence usage through system design and implementation. Venkatesh et al. stressed the need to extend the TAM literature by explicitly considering system and information characteristics and the way in which they might indirectly influence system usage.

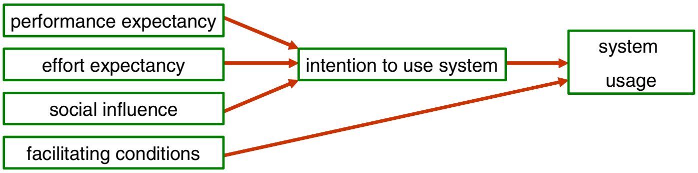


Fig. 1. Venkatesh's model [25]

3. Satisfaction: Attitude toward the System

In contrast to TAM, system and information characteristics have been core elements in the literature on user satisfaction. The DeLone study [7] is one of the first attempts at a comprehensive review of the literature on IS success. It organized a broad base of diverse research (180 articles) and presented a more integrated view of IS success. DeLone's model consists of six interdependent constructs for IS success: system quality (SQ), information quality (IQ), use, user satisfaction, individual impact, and organization impact (see Fig. 2). It identified IQ and SQ as antecedents of user satisfaction and use.

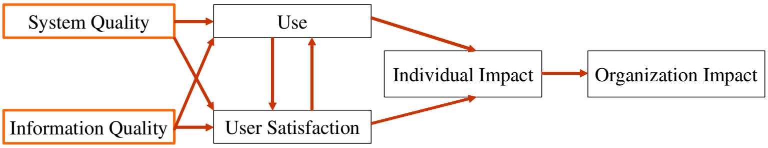


Fig. 2. DeLone's IS success model [7]

Seddon suggested that DeLone et al. tried to do too much with their model; as a result, the model is confusing and lacks specificity [22]. Seddon's major contribution is a re-specified model of IS success. Seddon defined success as a measure of the degree to which the person evaluating the system believes that the stakeholder is better off. The model shows that both perceived usefulness and user satisfaction depend on IQ, SQ, and benefits (see Fig. 3). Both DeLone and Seddon made an explicit distinction between information aspects and system features as determinants of user satisfaction.

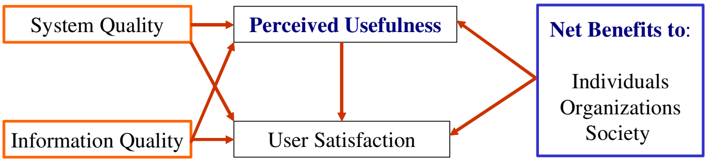


Fig. 3. Seddon's IS success model [22]

4. Information-seeking Behavior: Identify Temporal Users' Information Needs

Satisfaction is a consequence of the user's experience during various information-seeking stages. The changing needs of users interacting with the DL should be identified. Therefore, understanding of users' information-seeking behavior is required.

The information-seeking behavior of academic scholars has been studied for decades, and many models have been generated. Among them are Ellis's model [8] and Kuhlthau's model [16]. These two models are based on empirical research and have been tested in subsequent studies. Ellis's model includes six generic features coded from E1 through E6 as shown in Fig. 4. As of 2002, there were more than 150 papers that cite Ellis's information-seeking behavior model of social scientists [20]. Most of the information-seeking behavior features in Ellis's model are now being supported by capabilities available in Web browsers. Kuhlthau's model complements that of Ellis by attaching to stages of the information-seeking process the associated feelings, thoughts and actions, and the appropriate information tasks. The stages of Kuhlthau's model are coded from K1 through K6 as shown in Fig. 4. Kuhlthau's model is more general than that of Ellis in drawing attention to the feelings associated with the various stages and activities. It also has been applied to support learning from DLs [19].

3 DL Success Model

We further connect Gonçalves' DL quality model and the information life cycle model [5] with Ellis' and Kuhlthau's information-seeking behavior models as shown in Fig. 4. The outer arrows in Fig. 4 indicate the life cycle stage (active, semi-active, and inactive) for a given type of information. The innermost portion of the cycle has four major phases of information use or process: information creation, distribution, seeking, and utilization. Each major phase is connected to a number of activities.

Gonçalves stated that his work took a very system-oriented view of the quality problem and partially neglected its usage dimension. Our goal is to define the success of DL from an end user perspective; hence we focus on the 'seeking' and 'utilization' stages. Behaviors occurring at the 'seeking' phase and 'utilization' phase are elaborated in Fig. 4 by Ellis' and Kuhlthau's models. Each dimension of quality is associated with a corresponding set of activities. Quality dimensions associated with the seeking and utilization phases are related to constructs of the DL success model.

Our proposed DL success model consists of four interrelated and interdependent constructs based on the previously discussed theoretical methods. The general proposition of our model is that DL satisfaction and the intention to (re)use a DL are dependent on four constructs: information quality, system quality, performance expectancy, and social influence (see Fig. 5). Arrows in Fig. 5 indicate that a construct is affected by each construct that points to it. IQ and SQ can be found in the IS success literature, while performance expectancy and social influence can be found in the IT adoption literature. Since our model incorporates TAM, it is a predictive model, i.e., it can be used to predict intention to (re)use. We think determinants of success are goal and user specific. Hence, a measurement instrument of "overall success" based on arbitrary selection of items from the four constructs is likely to be problematic. Individual measures from the four constructs should therefore be combined systematically to create a comprehensive measurement instrument.

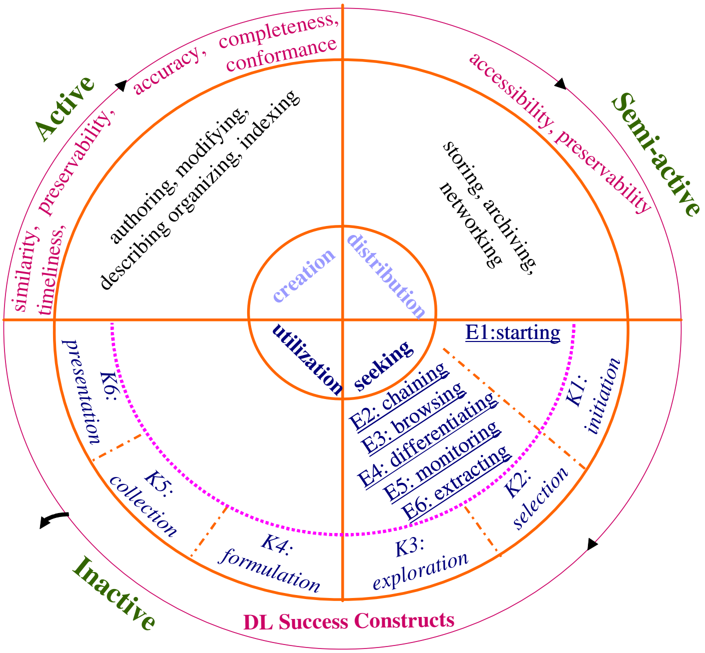


Fig. 4. Connection of DL quality model with information life cycle and information seeking behavior models

1. Information Quality (IQ)

Information in DLs can be classified from two different perspectives, the DL developers' view and the DL patrons' (end users') view. Five main concepts related to DL information within the 5S framework are: repository, collection, metadata catalog, digital object, and metadata specification (see Fig. 6). A DL repository involves a set of collections, each of which is a set of digital objects. Samples of digital objects can be electronic theses (or dissertations) and records of artifacts (such as bones, seeds, and figurines) excavated from an archaeological site. Each digital object is assigned associated metadata specification(s), which compose the metadata catalog.

While the dimensions of quality for each of the five concepts are defined in [11] and listed in the left part of Fig. 7, they do not fully differentiate end users from DL developers. We group the five concepts into three categories and develop six items (factors) to measure the quality for each of the three categories for end users, as shown in the right part of Fig. 7. The dashed arrows illustrate that parts of the quality dimensions discussed in [11] are associated with the six items measuring DL IQ.

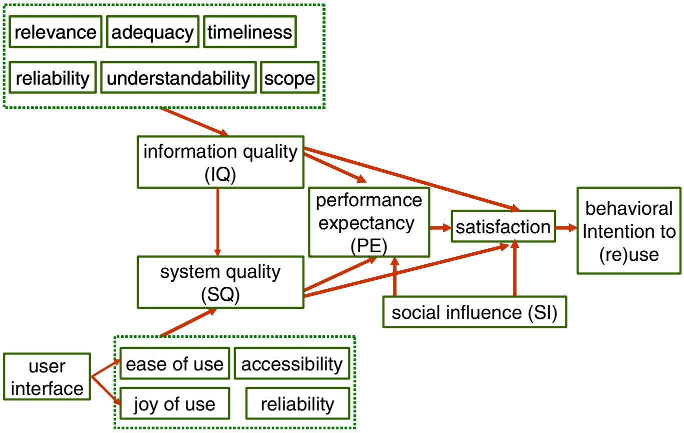


Fig. 5. DL success model (integrating Fig. 1- Fig. 3)

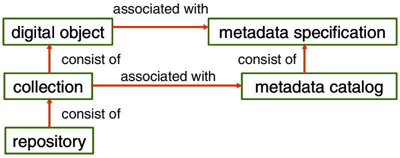


Fig. 6. Concepts related to DL information

a) Digital object and metadata specification:

Accuracy and completeness are defined in [11] as quality dimensions for metadata specifications, however, they are absent in the quality dimensions list for a digital object. This suggests two other quality measures for digital object and metadata specification: adequacy and reliability. Adequacy indicates the degree of sufficiency and completeness. Reliability indicates the degree of accuracy, credibility, and consistency.

Relevance is concerned with such issues as relevancy, pertinence, and the applicability of the information. Pertinence and relevance for digital objects are measured with Boolean values (0 or 1) in [11]. They are a subjective judgment by users in a particular context. We use relevance to measure the quality of both digital object and metadata specification. Significance of a digital object defined in [11] reflects relevance to user needs or particular user requirements. Therefore, significance can be partially mapped to relevance. Similarity metrics defined in [11] reflect the relatedness among digital objects. If one of the digital objects is a user's information need, then similarity is associated with the relevance item (factor).

Timeliness is concerned with the currency of the information. Understandability encompasses variables such as being clear in meaning and easy to understand.

Preservability as an important digital object quality property needs to be identified by DL developers; however, it may not be visible to DL patrons. The accessibility of a digital object is managed by DL services, so it is used to measure DL services instead of information. Therefore, preservability and accessibility are not included in the six items for DL IQ that are shown in Fig 7.

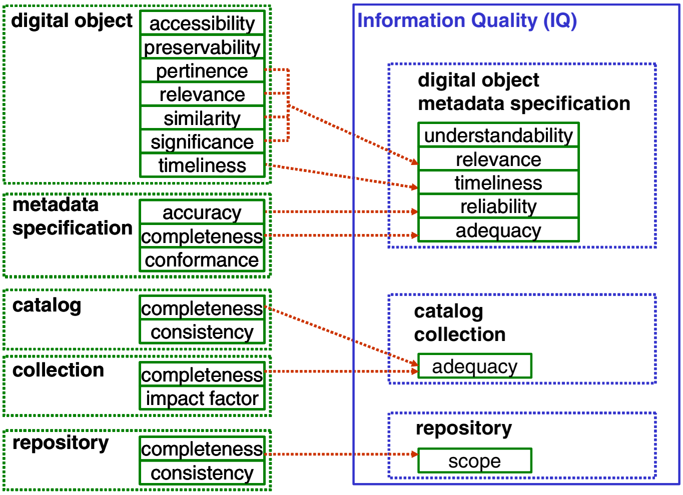


Fig. 7. DL information quality (IQ) measurement

b) Metadata catalog and collection

Adequacy is used to measure the degree of sufficiency and completeness of DL metadata catalogs and collections.

c) Repository

Scope evaluates the extent and range of the repository. These address the breadth of information and the number of different subjects. According to [11], a repository is complete if it contains all collections it should have. Therefore, completeness defined in [11] is associated with scope.

2. System Quality (SQ)

Dimensions of quality for DL services are classified as internal (e.g., top three entries) or external (e.g., bottom three entries) in [11], as shown in the dashed box in Fig. 8. We focus on the external view, concerned with the use and perceived value of these services from the end users' point of view. They relate to DL system quality (SQ) and performance expectancy (discussed in Section 3.3) as indicated by the three dashed arrows in Fig. 8. We develop four items to measure DL SQ.

Prior research subscales for accessibility include system responsiveness and loading time. The accessibility of a DL refers to not only its speed of access and availability but also to its **information** (e.g., **digital objects and metadata accessibility**). Efficiency defined in [11] is measured in terms of speed; it is associated with service accessibility. A DL needs to be reliable, which means that it is operationally stable.

Ease of use is concerned with how simple it is for users to (learn to) use DLs. Joy of use is about the degree of user pleasure. These two items are affected by the user interface through navigation and screen design as indicated by the two solid arrows

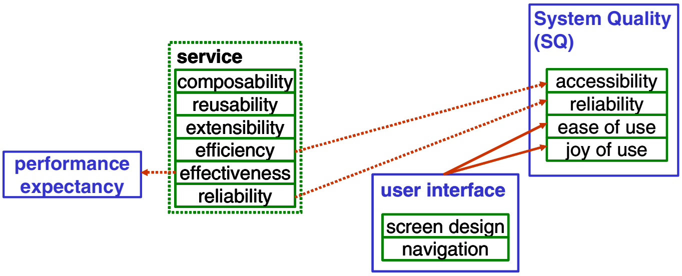


Fig. 8. DL service quality (SQ) measurement

shown in Fig. 8. Navigation is concerned with evaluating the links to needed information that are provided on the various pages of a DL website. Screen design is the way information is presented on the screen. It affects both ease of use and joy of use. Having an organized and well-designed screen aids users in locating relevant information more easily, while an attractive user interface helps increase joy of use. Although we have a common idea that aesthetic objects should be symmetric, balanced, or well proportioned, there is no general instruction set prescribing how to create aesthetic interfaces [12]

3. Performance Expectancy (PE)

Performance expectancy (see Fig. 5) is defined as the degree to which users believe that a specific DL will help them gain advantage in accomplishing their desired goal. In [25], it consists of five constructs: perceived usefulness, extrinsic motivation, job-fit, relative advantage, and outcome expectations.

4. Social Influence (SI)

Social influence (see Fig. 5) is concerned with a user's perception that other important people favor a particular DL. Many studies have been done in the marketing domain on the role of social influence. Accordingly, it seems appropriate to consider social influence on DL usage. As reported in [24], DL visibility is considered as an important factor that may lead to greater user acceptance of DLs. Potential users may not be aware of the benefits of using the DL, or even its existence. Increasing DL visibility can help users perceive the DL as more useful, although it will not increase the functionality of a DL.

5 Case Study

As part of the requirements analysis for an archaeological DL, ETANA-DL [23], email interviews with 5 prestigious archaeologists, and face to face workplace interviews with 11 archaeologists (including 3 of the 5 interviewed by email) were conducted. Subsequent formative evaluation studies were carried out to improve system design. In this section, we associate the four constructs of the model discussed in the previous section with the activities occurring in the seeking and utilization phases (see the innermost portion of the cycle in Fig. 4) by analyzing the results of the interviews and the formative usability studies. These results are shown in Table 1 and may help distinguish issues that are generic across domains, from those that are domain specific.

Table 1. DL success constructs associated with seeking and utilization phases

DL success Construct	*seeking phrase*		*utilization phrase*		
	starting (E1/K1)	*selection exploration (E2-E6)/(K2-K3)*	*formulation (K4)*	*collection (K5)*	*presentation (K6)*
social influence	*DL visibility*				
information quality		*adequacy, scope*	*accuracy*		
system quality		*ease of use* *joy of use (interface)*	*accessibility*	*accessibility*	*accessibility*
performance expectancy		*usefulness (interface)*			

1. Seeking phase

- E1/K1

"starting" activity in Ellis' model ('initiation' stage in Kuhlthau's model) is usually at the beginning of information seeking. It may help one 'recognize' a need for information. Users' information needs may be initiated by a specific active task or condition, or by requirements identified passively.

Social influence, such as regarding DL visibility, is associated with this stage. Within the archaeological domain, awareness of DLs is poor. Methods to increase DL visibility include:

1) Publicize the existence of a DL: One archaeologist said that "… the turning point for the DL will be when someone has demonstrated in a print publication how ETANA-DL helped in their research …". Some recommended more international collaboration, e.g., some suggested that ETANA-DL may consider collaboration with JADIS (Jordanian Archaeological Data Information System) to increase its visibility. Since JADIS is one of the main Jordanian cultural resource management systems, connecting ETANA-DL with JADIS could allow basic survey and overall information on Jordanian archaeology to be combined with ETANA-DL's more in-depth coverage.

2) Provide a DL alert service (e.g., press alerts): Archaeologists may want alerts when new artifacts from others arise on their subjects of interests.

- (E2-E6)/(K2-K3)

These five feature activities in Ellis's model ('chaining', 'browsing', 'differentiating', 'monitoring', and 'extracting') occur in the 'selection' and 'exploration' stages in Kuhlthau's model. In the 'selection' stage, a general area for investigation is identified (located). The appropriate task at this point is to fix the general topic of exploration. Exploration has many cognitive requirements similar to browsing and search tasks.

IQ, SQ, and PE are associated with these stages. Regarding IQ, adequacy (degree of sufficiency and completeness) of DL collections and metadata catalogs and scope of DL repository should be considered. Some archaeologists pointed out: "Ideally, the system would include as many types of data as possible, from text summaries to photos, maps, and other visuals."

Regarding SQ and PE, interface plays a major role in influencing the usefulness, easy of use, and joy of use. The quality of the DL interface makes a significant contribution to a usable DL, and interface problems often are cited by non-users as a major reason for not using electronic information retrieval systems [9]. As a virtual intermediary between users and a DL, the interface is the door through which users access a DL. The interface characteristics (screen design and navigation) that affect DL usability include those commonly found in most web GUIs, as well as the ones specific to archaeological DLs.

1) Screen design: The way that information is arranged on the screen can influence the users' interaction with DLs beyond the effect of the information content. Some archaeologists suggested that "… the interface needs to be more visually stimulating … should allow to browse visual stacks of the digital library…". Another issue to be considered for screen design is the wording for labeling. In the archaeological domain, an example could be the terminology for periodization schemas. There are different periodization schemas based on political, historical, or cultural events. The archaeologists found it difficult to use a single "standard" periodization schema.

2) Navigation: The navigation should enable archaeologists to explore a DL without having to keep an auxiliary memory aid like a yellow pad at hand.

2. Utilization phase

Information management and utilization was not identified as a category in Ellis's study of social scientists. On the other hand, the last three stages in Kuhlthau's model involve organizing information into a coherent structure.

- K4

The formulation stage is identified as conceptually the most important step in the process [16]. Users focus on a more specific area within the topic and make sense of (or interpret) information in the light of their own needs. A guiding idea or theme emerges which is used to construct a story or narrative, or to test a hypothesis. This formulation also will guide the users in selecting appropriate information.

Research has considered the process of interpreting documents (e.g., reading and annotating them) rather than simply locating them [3]. Within the archaeological domain, archaeologists formulate a personal perspective or sense of meaning from the encountered information. However, they usually conduct interpretation offline. Access to primary data and data analysis services provided by DLs enable archaeologists to make interpretations online, if they change work habits. Alternatively, exporting of results to files or into special formats like for spreadsheets may be helpful to support subsequent offline management, processing, visualization, and reporting.

Some sample factors affecting formulation are as follows.

1) Information accuracy: Formulation is associated with verifying the accuracy of the information found. Archaeologists need reputable (trusted) information or information analysis to support interpretation.

2) Information accessibility: It defines how much effort (time) is required to find (locate) the information needed. In the archaeological domain, primary data usually is available to researchers outside a project (site) only after substantial delay. Some archaeologists said that "… ETANA-DL would be a very efficient way to disseminate and share our research, and in turn, we could utilize the work of others as much as possible."

- K5

In the collection stage, information is gathered to support the chosen focus. Information accessibility is very important as discussed above.

- K6

During this final stage, presentation, ideas, focus, and collected resources are organized for publishing and sharing. Some archaeologist suggested making arrangement with the publishers of obscure journals to include their publications in ETANA-DL. They found that it is useful for ETANA-DL to provide a discussion forum to share their interpretation of annotated items.

6 Conclusions

The goals and objectives of a DL differ depending on the DL type, resulting in varying ideas of satisfaction as well as success. Therefore, to determine success across DLs from the perspective of users is goal and context specific. The work presented in this paper lays the foundation for defining success of DLs from the view of DL end users. Our work assumes a multi-theoretical perspective and synthesizes many related research areas in terms of theory and empirical work. Our case study illustrates and further explicates the approach, which we have shown to be helpful with regard to a DL to support Near Eastern archaeology. We will empirically validate the proposed model further when we apply it in various domain specific DLs in the future.

Acknowledgements. This work was funded in part by the National Science Foundation (ITR-0325579). We thank Marshall Breeding, Linda Cantara, Douglas Clark, Joanne Eustis, James W. Flanagan, and all the interviewees for their support. We also thank all our colleagues in the Digital Library Research Laboratory at Virginia Tech.

References

1. Adams, A. and Blandford, A. Digital libraries' support for the user's 'information journey'. In Proc. JCDL 2005, June 7-11, 2005, Denver, 160-169.
2. Agosti, M., Nunzio, G.M.D. and Ferro, N. Evaluation of a digital library system. In Proc. of the DELOS WP7 Workshop on the Evaluation of Digital Libraries, Padova, Italy, October 4-5, 2004, 73–76.
3. Bishop, A.P. Making Digital Libraries Go: Comparing Use Across Genres. ACM DL 1999: 94-103.
4. Blandford, A. and Buchanan, G. Usability of digital libraries: a source of creative tensions with technical developments. In IEEE-CS Technical Committee on Digital Libraries' online newsletter, Vol. 1, No. 1
5. Borgman, C.L. Social aspects of digital libraries. In DL'96: Proceedings of the 1st ACM International Conference on Digital Libraries, D-Lib Working Session, http://is.gseis.ucla.edu/research/dl/UCLA_DL_Report.html, 1996.
6. Champeny, L., Borgman, C.L., Leazer, G.H., Gilliland-Swetland, A.J., Millwood, K.A., D'Avolio, L., Finley, J.R., Smart, L.J., Mautone, P.D., Mayer, R.E. and Johnson, R.A. Developing a digital learning environment: an evaluation of design and implementation processes. In Proc. JCDL 2004: 37-46.
7. DeLone, W.H. and McLean, E.R. Information systems success: The quest for the dependent variable. Information Systems Research, 3 (1). 60-95, 1992.

8. Ellis, D. and Haugan, M. Modeling the information seeking patterns of engineers and research scientists in an industrial environment. Journal of Documentation, 53(4): 384-403, 1997.
9. Fox, E.A., Hix, D., Nowell, L.T., Brueni, D.J., Wake, W.C., Heath, L.S. and Rao, D. Users, User Interfaces, and Objects: Envision, a Digital Library. JASIS 44(8): 480-491 (1993).
10. Fuhr, N., Hansen, P., Mabe, M., Micsik, A. and Sølvberg, I. Digital libraries: A generic classification and evaluation scheme. Springer Lecture Notes in Computer Science, 2163:187–199, 2001.
11. Gonçalves, M.A. Stream, Structure, Space, Scenarios, and Societies (5S): A Formal Digital Library Framework and Its Applications, Ph.D. Dissertation, Virginia Tech, http://scholar.lib.vt.edu/theses/available/etd-12052004-135923, 2004.
12. Grün, C., Gerken, J., Jetter, H.-C., König, W. and Reiterer, H. MedioVis - A User-Centred Library Metadata Browser. In Proc. ECDL 2005: 174-185.
13. Hartson, H.R., Shivakumar, P. and Pérez-Quiñones, M.A. Usability inspection of digital libraries: a case study. Int. J. on Digital Libraries 4(2): 108-123 (2004).
14. Hill, L.L., Carver, L., Larsgaard, M., Dolin, R., Smith, T.R., Frew, J. and Rae, M.-A. Alexandria digital library: user evaluation studies and system design. JASIS 51(3): 246-259 (2000).
15. Kengeri, R., Seals, C.D., Harley, H.D., Reddy, H.P. and Fox, E.A. Usability Study of Digital Libraries: ACM, IEEE-CS, NCSTRL, NDLTD. Int. J. on Digital Libraries 2(2-3): 157-169 (1999).
16. Kuhlthau, C.C. Learning in digital libraries: an information search process approach. Library Trends 45(4): 708-724, 1997.
17. Larsen, R.L. and Wactlar, H.D. Knowledge Lost in Information: Report of the NSF Workshop on Research Directions for Digital Libraries, June 15-17, 2003, Chatham, MA, National Science Foundation Award No. IIS-0331314. http://www.sis.pitt.edu/~dlwkshop/.
18. Marchionini, G. Evaluating Digital Libraries: A Longitudinal and Multifaceted View preprint from Library Trends, 49(2): 304-333, 2000.
19. Marshall, B., Zhang, Y., Chen, H., Lally, A.M., Shen, R., Fox, E.A. and Cassel, L.N. Convergence of Knowledge Management and E-Learning: The GetSmart Experience. In Proceedings JCDL2003, Houston, 135-146, 2003.
20. Meho, L.I. and Tibbo, H.R. Modeling the information-seeking behavior of social scientists: Ellis's study revisited. JASIST 54(6): 570-587, 2003.
21. Saracevic, T. Digital library evaluation: Toward evolution of concepts. Library Trends, 49(2): 350–369, 2000.
22. Seddon, P.B. A respecification and extension of the DeLone and McLean model of IS success. Information Systems Research, 8 (3): 240-253, 1997.
23. Shen, R., Gonçalves, M.A., Fan, W. and Fox, E.A. Requirements Gathering and Modeling of Domain-Specific Digital Libraries with the 5S Framework: An Archaeological Case Study with ETANA. In Proceedings ECDL2005, Vienna, Sept. 18-23.
24. Thong, J.Y.L., Hong, W. and Tam, K.Y. What leads to acceptance of digital libraries? Commun. ACM 47(11): 78-83 (2004).
25. Venkatesh, V., Morris, M., Davis, G. and Davis, F. User acceptance of information technology: Toward a unified view. MIS Quarterly, 27 (3): 425-478, 2003.

Evaluation of Metadata Standards in the Context of Digital Audio-Visual Libraries

Robbie De Sutter, Stijn Notebaert, and Rik Van de Walle

Ghent University - IBBT, ELIS, Multimedia Lab,
Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium
robbie.desutter@UGent.be

Abstract. Digital file-based libraries for the audio-visual material of television broadcasters and production houses are becoming desirable. These libraries not only address the problem of loss of content due to tape deterioration, but also improve disclosure of the content. However, switching to a digital file-based library involves many new concerns and problems for content providers. This paper will discuss one of them, namely the metadata. Metadata is additional information that is required in order to be able to search, retrieve, and play out the stored content. Different standards for metadata are currently available, each having its own field of application and characteristics. In this paper, we introduce an objective framework that one can use in order to select the appropriate metadata standard for its particular type of application. This framework is applied to four well-known metadata standards, namely Dublin Core, MPEG-7, P/Meta, and SMEF.

1 Introduction

Digital file-based libraries with audio-visual material originating from television broadcasters and television production houses are still uncommon. Most of them store and archive their audio-visual material on tape. Unfortunately, these tapes deteriorate over time, resulting in content loss. Broadcasters recognize this problem and plan to switch to a tapeless archive. This switch is feasible as the price to store audio-visual material as digital files on hard disks is acceptable. Furthermore, a tapeless file-based archive improves the disclosure of the material.

The switch to a tapeless archive encounters similar problems as those the librarians encountered during the digitalization of their libraries, such as the need for particular *metadata* — i.e., additional information about the material. In [1] it is observed that "the metadata necessary for successful management and use of digital objects is both more extensive than and different from the metadata used for managing collections of physical material." This statement holds true for audio-visual material. For example, the appropriate technical metadata must be available to be able to play out the material. Furthermore, it is also the metadata that will ensure that the material in the archive can be searched and retrieved.

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 220–231, 2006.

The *Society of Motion Pictures and Television Engineers* (SMPTE)[1] uses the following definition:

$$content = essence + metadata$$

Here, the *essence* is the audio-visual material. The definition states that without metadata, there is no content: content cannot be found or used without metadata; hence the essence is unusable. A second definition extends the previous definition:

$$asset = content + right\ to\ use\ it$$

This definition states that content is only valuable (i.e., an *asset*) if the content owner has the right to utilize it.

It is clear that, in order to have a fully functional and usable digital audio-visual library, choosing the best suited metadata standard is the key to success. Indeed, more and more new metadata standards with as main purpose to annotate and manage audio-visual material are available. However, as these standards are intended for different fields of application, selecting the "best" standard depends on the intended use.

In this paper, we define different criteria that can be used to evaluate and compare metadata standards. As such, these criteria allow one to make a well-considered choice. Furthermore, we apply the criteria to four well-known metadata standards, namely Dublin Core [2], the Multimedia Content Description Interface (also known as MPEG-7) [3, 4], P/Meta [5], and the Standard Media Exchange Framework (SMEF) [6].

The remainder of the paper is organized as follows. In section 2, we give an overview of the related work. Next, in section 3 we define the different evaluation and comparison criteria for metadata standards. Subsequently in section 4, these criteria are applied to the four metadata standards. Finally, section 5 concludes this paper.

2 Related Work

The expertise built up by the librarians when creating digital libraries is of great value for any other digitization effort, also for the digitalization of the audio-visual archives of television broadcasters and archives alike. The purpose of a digital library – as seen by the librarians – is described in [7] as "electronic libraries in which large numbers of geographically distributed users can access the contents of large and diverse repositories of electronic objects – networked text, images, maps, sounds, videos, catalogues of merchandize, scientific, business and government data sets – they also include hypertext, hypermedia and multimedia compositions."

This statement emphasizes that the library community mainly focuses on the disclosure and the exchange of digital objects. This resulted in the creation of the *Metadata Encoding & Transmission Standard* (METS) by the Library of

[1] More information on the Society of Motion Pictures and Television Engineers can be found at `http://www.smpte.org`

Congress [8]. METS provides a format for encoding the metadata used for the management and the exchange of digital objects stored within the library, and this by extending the techniques developed by *the making of America II (MOA II)* project [9]. However these standards do not normatively fix the structural, administrative, and technical metadata itself. Furthermore they only refer to available techniques in the pre-digital libraries community for descriptive metadata, such as *Machine-Readable Cataloging* (MARC) records [10] or the *Encoded Archival Description* (EAD) [11]. These pre-digital documentation techniques are inadequate to fully document digital works. Better suitable standards for the digital libraries are more and more being investigated and used, such as the *Dublin Core Metadata Element Set*, the *General International Standard Archival Description* [12] and the *Multimedia Content Description Framework* (also known as MPEG-7) [13]. For an overview of these standards, we refer the reader to [14].

Most efforts and investigations in the digital library communities are about ways to exchange digital objects between repositories and to ensure interoperability thereof. The *Open Archive Initiative Protocol for Metadata Harvesting* (OAI-PMH) is a major effort to address technical interoperability among distributed archives [15, 16, 17]. This initiative lays down the fact that the *Simple Dublin Core Metadata Element Set* as defined by the Dublin Core Metadata Initiative is the baseline to assure metadata interoperability.

Unfortunately, the used techniques for creating and maintaining digital libraries of books and images – based on METS and the results of the MOA2 project – "lacks adequate provisions for encoding of descriptive metadata, only supports technical metadata for a narrow range of text- and image-based resources, provides no support for audio, video, and other time dependent media, and provides only very minimal internal and external linking facilities" [1]. This implies that solely using these technologies to create audio-visual digital libraries is insufficient. These concerns are addressed by new metadata standards with as main purpose to annotate and manage audio-visual material, such as P/Meta and SMEF. The remainder of the paper focuses on a framework to compare the metadata standards intended for annotation of audio-visual libraries.

3 Selection Criteria

This section describes criteria that one can use to select the metadata standard that is best suited for the application in mind. These criteria are composed in such a way that all aspects ranging from content organization to the different types of metadata are taken into account, but independent to any restriction imposed by a particular media asset management system.

3.1 Criterion 1: Internal vs. Exchange Metadata Model

For this first criterion, it is important to identify the involved parties that exchange audio-visual material during their typical life cycle. The *European Broadcasting Union* (EBU) identifies in [18] the consumers and three trading entities, being the content creator, the content distributor, and the archive. EBU has investigated the different relationships between these four players and has presented the entities and the relationships in the EBU P/Meta Business-to-Business Dataflow Model (see Fig. 1). This model is independent of any metadata model and is applicable to most broadcasters.

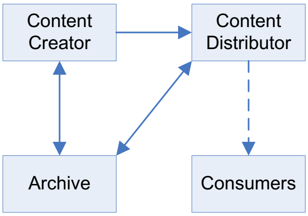


Fig. 1. EBU P/Meta Business-to-Business Dataflow Model [18]

On the one hand, particular metadata models are specifically developed for managing the metadata in the interior of a system. These metadata models are further referred to as *internal metadata models*. Usually, these metadata models are represented as *Entity Relationship Diagrams* (ERDs) which describe the architecture of the database that stores the metadata of the audio-visual material.

On the other hand, other metadata models are used to describe the way the information is to be transmitted from source to destination. Here, the metadata models are called *exchange metadata models*. These models are used to exchange information about the audio-visual material and are specifically intended for the transmission of metadata between different systems. Here, exchange must be seen as broad as possible, namely between any combination of content creator, content distributor, archive, and consumers.

3.2 Criterion 2: Flat vs. Hierarchical Metadata Model

The structural organization of the description of the essence is a second criterion. In general, the broadcaster decides how detailed the metadata needs to be. Two extreme visions can be identified. On the one hand the essence is considered as an elementary and indivisible unit, resulting in a coarse description, and on the other hand the essence is divided in small sub-pieces each annotated separately, resulting in a fine-detailed description.

If the essence is considered as an elementary and indivisible unit, the broadcaster can associate this elementary unit with, for example, a *program*. The metadata describes the essence (here this is that program) as a whole and does not describe the individual parts therein. This model is mostly referred to as a *flat metadata model*.

Sub-parts of the essence can be annotated with much more detail. The additional metadata belongs to the individual parts and permits the users of the archive to perform more detailed searches on the content. For example, a program can be split up in several *editorial objects*, corresponding to, for example, the individual scenes. Every editorial object can be annotated with additional descriptive metadata, so it is possible to search on the editorial object itself. In turn, editorial objects can be broken down in different *media objects*. These media objects could be, for example, the audio components, the video components, the subtitles, and so on. This is also possible the other way around: a group of programs which belong together can be collected in a *program group*. This program group is annotated with information identical to all programs within, for example, the name of the program. Hence, it is not necessary to repeat the same information for every program, but the program inherits information from its program group. The underlying idea is that information has to be added to the objects at the right location. This concept is referred to as a *hierarchical metadata model*. A four-layered architecture as discussed above is visualized in Fig. 2.

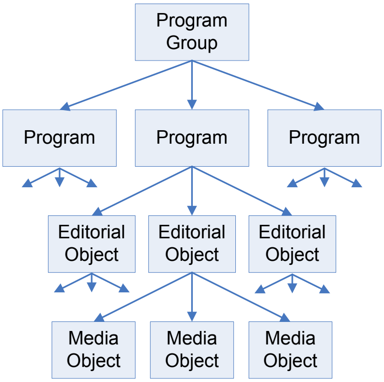


Fig. 2. Content organization

The broadcaster will not always want to use a hierarchical metadata model although this has huge benefits for faster and more efficient search and retrieve operations. Indeed, the most important reason for a broadcaster to restrict the metadata (and thus the decomposition of the audio-visual asset), is to limit the increase of cost which is proportionally to the amount of metadata that needs to be collected. It is clear that, as the metadata about an audio-visual object grows, the marginal profit of the additional metadata decreases, but the cost to generate this additional metadata increases disproportional. At a certain moment, it will be impossible to add additional metadata without making unjustified costs. In other words, the broadcaster will have to make a trade-off between cost and comprehensiveness of the metadata.

3.3 Criterion 3: Supported Types of Metadata

Metadata describes the essence. The requirements of the users determine the needed types of metadata. There are two rules that must be observed as explained in the introduction of this paper: 1) essence is unusable withoutmetadata, and 2) the content is valueless without rights information. Hereafter different types of metadata for the preservation of audio-visual material are discussed.

Identification metadata. The identification metadata is primarily about the information to singularly identify the essence. This can be done by human interpretable fields, like a title or an index, or by machine understandable identifiers, like a *Unique Material Identifier* (UMID) or a *Uniform Resource Identifier* (URI). Besides the identification metadata related to the essence, other identifying information is necessary to locate related documents that are potentially stored in another system.

Description & classification metadata. The descriptive metadata must describe what the essence expresses. This could be done by providing a list of keywords which try to place the essence in a particular context. In some cases, the keywords are selected from an organized dictionary of terms, i.e., the *thesaurus*. Other than the concept of the thesaurus for purely descriptive purposes, other classification schemes can be used. Indeed, the content can be categorized in different predefined classes in accordance with the genre, the audience, and so on. A very well-known classification system is the Escort 2.4 system [19] from the EBU that groups the essence in conceptual, administrative, productional, scheduling, transmission, viewing and financial ways.

Another type of descriptive metadata comprises the description of the essence as a short text. This type of descriptive metadata is well-known and therefore it will be extensively used in practically every archiving system. Unfortunately, these fields are error-prone (e.g., spelling mistakes) and should be used carefully.

Technical metadata. The technical metadata describes the technological characteristics of the related essence. The minimal required technical metadata must specify the audio and video codecs that can be used for the decoding of the audio-visual material. With this minimal information, the user has the possibility to play out the essence. Hence, the technical metadata enables the essence to become usable which is one of the key requirements in order to create content.

Security & rights metadata. The security metadata handles all aspects from secure transmission (i.e., the encryption method) to access rights. The latter augments the content into an asset. The access rights metadata can be split up in information about the rights holder and information about contracts. The rights holder is the organization who owns the rights of the audio-visual material. Also the contracts related to the publication of the content and the contracts of the people who are involved with the creation of the essence, are considered as rights metadata.

Publication metadata. The last type of metadata describes the publication(s) of the essence. Every publication establishes a date of publication, the time and the duration of the publication, the channel of publication, and so on. That way the broadcasters have an idea on the frequency and the popularity of the essence. Furthermore, this information is important to clear broadcasting rights and handle payments.

3.4 Criterion 4: Syntax and Semantics

Some standards define only syntax, others only semantics, and some define both. The syntax defines how the representation of the metadata must be done. One of the most important questions about the syntax is the choice between a textual and a binary representation. The textual representation has the advantage that the metadata is human readable, but at the same time it is very verbose. The binary representation is dense, but it has the disadvantage that it can only be handled by machines.

In case of plain text notation, the Extensible Markup Language (XML) is mostly used. If so, the metadata standard provides, besides the standard itself, an XML Schema that punctiliously determines the syntax of the metadata. Using the XML Schema makes it possible to check the correctness (i.e., *validity*) of the metadata. This characteristic enables interoperability.

The semantics of the metadata standard determine the meaning of the metadata elements. Without any semantic description, one is free to assume the denotation of the different metadata elements, presumably resulting in different interpretations thereof between users. Only if the description of the metadata elements is *closed* (i.e., every metadata element is semantically described), all users must agree on the sense of the metadata elements improving the interoperability.

4 Evaluate Metadata Standards

In this section, we apply the evaluation criteria of Section 3 to four well-known metadata standards, namely Dublin Core, MPEG-7, P/Meta, and SMEF. Table 1 gives an overview of the evaluation criteria for the four metadata standards.

4.1 Dublin Core

The *Dublin Core Metadata Initiative* (DCMI)[2] is an open consortium engaged in the development of interoperable and online metadata standards that support a broad range of purposes and business models. The DCMI defined in 1999 the *Simple Dublin Core Metadata Element Set* consisting of 15 elements. In a second phase, the model was extended with three additional elements and a series of refinements, resulting in the *Qualified Dublin Core Metadata Element Set* (DCMES) specification [2].

The goal of the DCMES specification is to exchange resource descriptions aiming at cross-domain applications (*criterion 1*). While both the Simple and the Qualified specifications are very straightforward, they suffer from two very important shortcomings. On the one hand, there are no provisions for describing hierarchically structured audio-visual content – however, this can be circumvented by making implicit references to other parts – hence DCMES is a flat metadata model (*criterion 2*). On the other hand, the number of available metadata elements is too limited for thoroughly annotating audio-visual resources in

[2] More information on DCMI can be found at `http://dublincore.org`

digital libraries. In particular, due to the generic character of DCMES, the metadata for describing the technical, rights, security, and publication information is very confined (*criterion 3*).

With regards to *criterion 4*, the semantics are concisely described, still the user has considerable freedom for own interpretation. The DCMI provides different ways for syntactically describing the metadata: there are guidelines for incorporating Dublin Core in XML[3] and guidelines to use Dublin Core in combination with the Resource Description Framework[4].

4.2 MPEG-7: Multimedia Content Description Interface

The *International Organization for Standardization* and the *International Electrotechnical Commission* (ISO/IEC) have created the International Standard 15938, formally named *Multimedia Content Description Interface*, but better known as the MPEG-7 standard, which provides a rich set of tools for thoroughly describing multimedia content [3, 4].

The MPEG-7 standard is developed for, among other things, the exchange of metadata describing audio-visual content. It has been designed to support a broad range of applications, without targeting a specific application. As such, it is an exchange model (*criterion 1*).

The MPEG-7 standard normatively defines the syntax, by using an XML Schema, and the semantics, via normative text, of all metadata elements (*criterion 4*). The elements are structured as *descriptors* and *description schemes*: a descriptor is defined for the representation of a particular feature of the audio-visual content, a description scheme is an ordered structure of both descriptors and other description schemes. This system is used to create a hierarchical model (*criterion 2*). For example, an audio-visual material can be described by its temporal decomposition and by its media source decomposition. The latter is divided into descriptions about the audio segment and the video segment, which is on its turn decomposed into shots, key frames, and objects.

The supported types of metadata (*criterion 3*) are mostly focused on the description, technical, and, to a lesser degree, identification metadata. Almost no attention was paid to publication and rights & security metadata elements, however ISO/IEC address these concerns in different parts of the MPEG-21 standard.

Part three and part four of the MPEG-7 standard handle about the technical metadata for video respectively audio content. In part five of the MPEG-7 standard, sometimes referred to as *Multimedia Description Schemes* (MDS), the descriptors and the description schemes for the description and classification of audio-visual material are defined. More information about MDS is given in [20], and an overview of the different functional areas is visualized in Fig. 3.

[3] The guidelines for the notation of Dublin Core in XML format can be found at http://dublincore.org/documents/dc-xml-guidelines

[4] More information on using Dublin Core in combination with Resource Description Framework can be found at http://dublincore.org/documents/dcmes-xml and http://dublincore.org/documents/dcq-rdf-xml

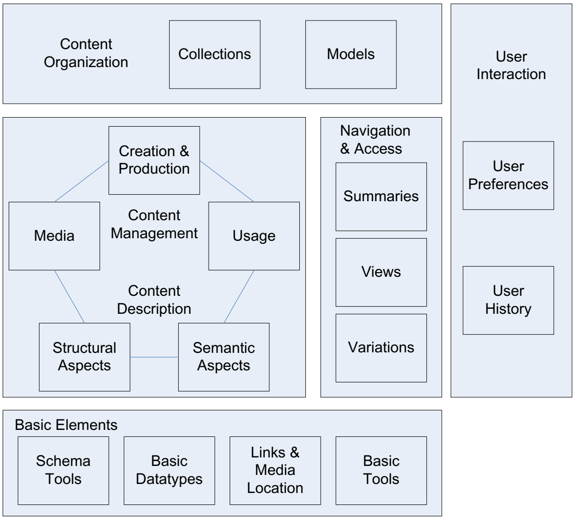


Fig. 3. Overview of Multimedia Description Schemes [20]

4.3 P/Meta

The P/Meta standard is developed by the EBU as a metadata vocabulary for program exchange in the professional broadcast industry [5]. Hence, it is not intended as an internal representation but as an exchange format for program-related information in a business-to-business use case (*criterion 1*).

The P/Meta standard [21] presents a five-layered hierarchical model (*criterion 2*): the brand, the program group, the program, the program item, and the media object. A brand collects all the program groups with a recognizable collective identity, e.g. information about the broadcasting TV-station. Every program group is composed of individual programs, which consist of individual program items. Finally every program item may be split up in media objects. This hierarchy is comparable with the one illustrated in Fig 2.

To obtain this hierarchical structure, the standard defines a number of sets and attributes. A P/Meta set groups P/Meta attributes and other P/Meta sets in such a way that all relevant metadata is collected for describing the considered object. A program group with the corresponding programs is described. Every program group and every program is annotated with identification (numbers and titles), classification (according to the Escort 2.4 system [19]), and description metadata. Besides these three elementary types, the description of the individual programs is complemented with four additional types, namely transmission or publication metadata, metadata concerning editorial objects and media objects, technical metadata (audio and video specification, compression schemes, and so on), and rights metadata (contract clauses, rights list, and copyright holders). These are also the supported types of metadata (*criterion 3*).

P/Meta defines all sets and attributes, resulting in a metadata standard where every term is determined unambiguously. The syntax is defined by an XML Schema (*criterion 4*).

4.4 Standard Media Exchange Framework

SMEF has been developed by the Media Data Group of BBC Technology, now Siemens SBS, on behalf of the British Broadcasting Corporation (BBC). Through a close collaboration with a wide variety of BBC projects, a profound understanding of the broadcaster's audio-visual media information requirements has been derived. Although the model is developed for use within the BBC, the definitions are organization independent and should be usable for any other broadcasters.

SMEF provides a rich set of data definitions for the range of information involved in the production, development, use, and management of media assets [6]. Its purpose is to ensure that different systems store this information in an equal way. Therefore, the SMEF standard defines an ERD which provides a framework for storing the metadata in the system. Hence, this is an internal metadata model (*criterion 1*).

The SMEF metadata model records all information that becomes available during the whole production cycle, from a program concept over media and editorial objects to the actual publication. An editorial object (i.e., the program) can be split up in different media objects, making this a hierarchical metadata model (*criterion 2*). Each media object can be annotated extensively with descriptive and technical metadata. The entities Editorial Object and Media Object can be linked with two other entities, namely the Usage Restriction entity (describing the restrictions on the use) and the Copyright Reporting entity (describing copyright details on the source material used). Hence, SMEF pays much attention to the rights metadata (*criterion 3*).

With regards to *criterion 4*, the SMEF standard defines the semantics of all entities, attributes, and relationships. The definition of syntactical rules covers the way the metadata is represented in the internal system.

Table 1. Overview of the Evaluation of the Metadata Standards

	Dublin Core	MPEG-7	P/Meta	SMEF
criterion 1	exchange	exchange	exchange	internal
criterion 2	flat	hierarchical	hierarchical	hierarchical
criterion 3	identification, description, and technical (limited)	identification, description, and technical	all	all (extensive rights metadata)
criterion 4	XML & RDF (a) open semantics	XML closed semantics	XML closed semantics	ERD open semantics

(a) DCMES can be mapped to XML and RDF.

5 Conclusions

In this paper, we discussed the need for digital file-based libraries for the audio-visual materials of television broadcasters and production houses. It is indispensable that these audio-visual materials are described by additional information, i.e. metadata, such that the materials in the libraries can be disclosed. It is the metadata that augments the materials from essences over content to assets.

Different metadata models exist to do this, whereby each model is suitable for a specific type of application. Within the paper, we introduced different selection criteria that one can use to compare and to select an appropriate metadata model for his intended application. The four selection criteria are 1) internal versus exchange metadata model, 2) flat versus hierarchical metadata model, 3) the supported types of metadata, and 4) the syntax and semantics of the model. To conclude this paper, we briefly introduced four well-known metadata standards, namely Dublin Core, MPEG-7, P/Meta, and SMEF, and applied the evaluation criteria.

Acknowledgment

The research activities that have been described in this paper were funded by Ghent University, the Interdisciplinary Institute for Broadband Technology (IBBT), the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research-Flanders (FWO-Flanders), the Belgian Federal Science Policy Office (BFSPO), and the European Union.

References

1. McDonough, J., Proffitt, M., Smith, M.: Structural, technical, and administrative metadata standards. A discussion document. Technical report, Digital Library Federation (2000) Available at http://www.diglib.org/standards/stamdframe.htm.
2. Dublin Core Metadata Initiative: Dublin core metadata element set, version 1.1: Reference description. Technical report (2004) Available at http://www.dublincore.org/documents/dces/.
3. Martínez, J.M., Koenen, R., Pereira, F.: MPEG-7: The Generic Multimedia Content Description Standard, Part 1. IEEE MultiMedia **9** (2002) 78–87
4. Martínez, J.M.: MPEG-7: Overview of MPEG-7 Description Tools, Part 2. IEEE MultiMedia **9** (2002) 83–93
5. Hopper, R.: Metadata exchange standards. Technical Report Technical Report No. 284, European Broadcasting Union (2000) Available at http://www.ebu.ch/en/technical/trev/trev_284-hopper.pdf.
6. BBC Technology: SMEF data model version 1.10. (Technical report) Available at http://www.bbc.co.uk/guidelines/smef/.
7. Sreenivasulu, V.: The Role of a Digital Librarian in the Management of Digital Information Systems (dis). Aslib Proceeding **18** (2000) 12–20
8. Digital Library Federation: METS: Metadata encoding and transmission standard. Technical report (2005) Available at http://www.loc.gov/standards/mets.

9. Digital Library Federation: The making of America II. Technical report (2005) Available at http://sunsite.berkeley.edu/MOA2.
10. Library of Congress: Understanding Marc Authority Records. Cataloging Distribution Service (2003)
11. The Society of American Archivists: Encoded Archival Description: Tag Library. Society of American Archivists (2002)
12. International Council on Archives: ISAD(G): General International Standard Archival Description, Second edition. (1999)
13. Manjunath, B.S., Salembier, P., Sikora, T.: Introduction to MPEG-7: Multimedia Content Description Language. John Wiley & Sons (2002)
14. Bekaert, J., Van De Ville, D., Strauven, I., De Kooning, E., Van de Walle, R.: Metadata-based Access to Multimedia Architectural and Historical Archive Collections: a Review. Aslib Proceeding **54** (2002) 362–371
15. Yu, S., Chen, H., Chang, H.: Building an Open Archive Union Catalog for Digital Archives. The Electronic Library **23** (2005) 410–418
16. Van de Sompel, H., Lagoze, C.: The Santa Fe Convention of the Open Archives Initiative. D-Lib Magazine **6** (2000)
17. Lagoze, C., Van de Sompel, H.: The making of the Open Archives Initiative Protocol for Metadata Harvesting. Library Hi Tech **21** (2003) 118–128
18. Hopper, R.: Metadata exchange scheme, v1.0. Technical Report Technical Report No. 290, European Broadcasting Union (2002) Available at http://www.ebu.ch/trev_290-hopper.pdf.
19. European Broadcasting Union: Escort: EBU System of Classification of RTV Programmes. Technical report (1995) Available at http://www.ebu.ch/en/technical/metadata/specifications.
20. Salembier, P., Smith, J.R.: MPEG-7 Multimedia Description Schemes. IEEE Transactions on Circuits, Systems and Video Technology **11** (2001) 748–759
21. European Broadcasting Union: P/Meta Metadata Exchange Scheme v1.1. Technical Report Tech. 3295 (2005) Available at http://www.ebu.ch/en/technical/metadata/specifications/notes_on_tech3295.php.

On the Problem of Identifying the Quality of Geographic Metadata[*]

Rafael Tolosana-Calasanz, José A. Álvarez-Robles, Javier Lacasta, Javier Nogueras-Iso, Pedro R. Muro-Medrano, and F. Javier Zarazaga-Soria

Computer Science and Systems Engineering Department,
University of Zaragoza
María de Luna, 1 50018 Zaragoza Spain
{rafaelt, jantonio, jlacasta, jnog, prmuro, javy}@unizar.es

Abstract. Geographic metadata quality is one of the most important aspects on the performance of Geographic Digital Libraries. After reviewing previous attempts outside the geographic domain, this paper presents early results from a series of experiments for the development of a quantitative method for quality assessment. The methodology is developed through two phases. Firstly, a list of geographic quality criteria is compiled from several experts of the area. Secondly, a statistical analysis (by developing a Principal Component Analysis) of a selection of geographic metadata record sets is performed in order to discover the features which correlate with good geographic metadata.

1 Introduction

Geographic Digital Libraries typically use geospatial metadata in order to provide surrogate representations of geographic resources and they represent the most powerful technique currently available for describing and locating geographic objects. As research and development make progress in the geographic area and metadata repositories grow in size (there are currently geospatial repository projects operating, whilst others are either to receive geographic metadata or plan to receive them in the near future), new requirements arise and system performance must improve necessarily. In this sense, the issues surrounding the creation of good quality metadata for Geographic Digital Libraries have surprisingly received little attention. Besides, regarding computer systems, there is a popular acronym, GIGO (Garbage In, Garbage Out), which means that if the input data is wrong, the output data will be unavoidably inaccurate or wrong.

[*] This work has been partially supported by the Spanish Ministry of Education and Science through the project TIC2003-09365-C02-01 from the National Plan for Scientific Research, Development and Technology Innovation. J. Lacasta's work has been partially supported by a grant (ref. B139/2003) from the Aragón's Government. Special thanks should be given to A. Sánchez, M. A. Manso, C. Fernández, P. Pachól, S. Fontano, A. Amaro and S. Muñoz.

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 232–243, 2006.

In other words, low quality information leads to bad system performance. Consequently, Geographic Digital Libraries need good quality metadata records in order to produce good results. The influence of poor quality metadata on the performance of Digital Libraries has been already studied from the perspective of other domains of knowledge: Barton [1] warned that "...these problems manifest themselves in various ways, including poor recall, poor precision, inconsistency of search results, ambiguities and so on...". Regarding the geographic domain, not only must the attention be focused on those problems, but also on the new ones that may appear with the geospatial information specific aspects: geographic coordinates, place names and so on.

Nevertheless, in order to tackle the problem, the requirements surrounding good quality metadata and, speaking more generally, the idea of quality have to be analysed previously. Quality is a matter of human judgement, thus, many complex human factors have a great influence on it. Additionally, it should be taken into consideration that these factors might vary widely among individuals or, what complicates things more, some individuals may modify their judgements throughout the time. However, the "notion" of quality is so simple, immediate and direct that it might be recognised less often by logical argument than by direct perception and observation. Mainly because of these reasons, much of the scientific research agrees that the definition of metadata quality is not out of difficulties. Nonetheless, according to [2] a metadata record of good quality is defined as "a record that is useful in a number of different contexts, both with respect to the search strategies and terms that can be used to locate it". Another definition [3], even more simple, might be "fitness for purpose". Following with this rationale, it seems that geographic metadata may be fit to their purpose, if they describe geographic data well and those descriptions are useful for their users.

The objective of this paper is to propose a quantitative method for quality assessment of metadata in geographic digital libraries. The method is developed through two phases, involving human experts in geographic information systems. Firstly, a list of geographic quality criteria, structural and semantic, is compiled from the experts. Then, derived from this criteria list, a group of metrics is proposed. Secondly, a statistical analysis of a selection of geographic metadata record sets is performed in order to discover the features which correlate significatively with good geographic metadata.

The remainder of this paper is organised as follows. Next section discusses other work related to this paper. In section 3, some geographic quality criteria are obtained from an opinion poll conducted to some experts and some geographic metadata metrics are proposed. In section 4, the statistical analysis is described and tested. Finally, the conclusions are given.

2 Related Work

Initial efforts in metadata development have been primarily invested in structure rather than in content, that is, in the design and in the implementation of geographic standards. Consequently, appropriate standards such as CSDGM [4] and

ISO19115 [5] were developed and currently represent an excellent base for metadata creation and system interoperability. However, not only does metadata quality depend on these standards, but also on the creation process. Thus, generally speaking, two main approaches can be found in the research of metadata quality.

On the one hand, some studies are more concerned with the content of the metadata fields and the process involved in the creation of the metadata. In [1] it is stated that once a metadata standard has been implemented within a system, the specified fields must be filled out with real data about real resources and this process brings its own problems. The following assumptions underlying the metadata creation process in the learning objects and the e-Prints communities are also challenged there:

- in the context of the culture of the Internet, mediation by controlling authorities is detrimental and undesirable, that rigorous metadata creation is too time-consuming and costly, a barrier in an area where the supposed benefits include savings in time, effort and cost.
- only authors and/or users of resources have the necessary knowledge or expertise to create metadata that will be meaningful to their colleges
- given a standard metadata structure, metadata content can be generated or resolved by machine.

Guy [3] suggests a number of quality assurance procedures that people setting up an e-Print archive can use to improve the quality of their metadata. The process is developed in the conviction that the metadata creation process is crucial to the establishment of a successful archive. Another interesting document is the report elaborated by the Academic ADL Co-Lab [2], which sets up the first step towards community creation and building in the learning repositories community. The paper is a guide to the various issues challenging learning repository projects: issues of quality, both content and metadata (creating quality content and metadata, guidelines to ensure access to quality educational content, quality and consistency of metadata, tools and workflow).

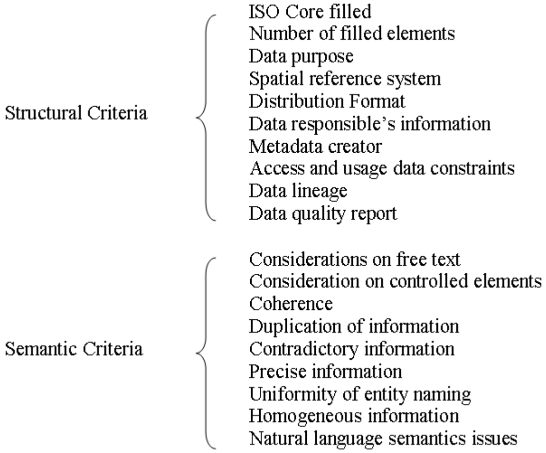
On the other hand, there exists another block of strategies whose research is mainly concerned with identifying and computing metrics for quality indicators. Then, resources are classified into different quality bands in accordance with those indicators. The study carried out by Armento [6] predicts quality rated Web documents (around popular entertainment topics) by using some preexisting relevance ranking algorithms. Armento states that the results, though promising, should be tested more extensively and with more quantity of data in other knowledge domains. Other experiments carried out by Custard and Summer [7] identify and compute metrics for sixteen quality indicators (indicators that were obtained from an extensive and previous literature review and meta-analysis) and employ machine-learning techniques in order to classify educational resources into different quality bands based on these indicators. Additionally, previous experiments were developed to determine whether these indicators could be actually used for the classification. Hughes [8] describes the motivation, design and implementation of an architecture to aid metadata quality assessment in the Open Archives Language (OAL) Community. It is worth highlighting that these quality indicators used in order to support quality judgements are based on the adherence to best practice guidelines for the use of the Dublin Core [9] elements and codes. Finally, another interesting work [10] computes some metrics for quality indicators and studies the relation between metadata quality and the quality of services.

3 Identifying Geographic Metadata Quality Criteria

At an early stage of our work, we considered studying the criteria by which the quality of geographic metadata records can be analysed. We carried out an initial experiment which consisted in asking several experts about the features, the elements or even the requirements for geographic metadata records that can determine their quality. As an outcome of this study, a compilation of geographic metadata quality criteria was obtained (see Fig. 1).

Two main tendencies can be observed in the compiled list. One tendency is more concerned with the structure of the metadata records and tries to determine to what extent the metadata records accomplish the standard. For instance, in the ISO 19115 standard, there exist certain recommendations regarding the format of certain data types such as dates, integers and so on. Additionally, in the same standard, there is a subset of elements known as the "ISO19115 Core metadata for geographic datasets" (called ISO Core onwards) that suggests to have each of them filled in. The most important elements such as the title, the abstract and the spatial reference system, among others, are included there. In the same sense, several experts were expecting to find specific information elements which were useful for their daily work and which were outside that core. Other considerations pointed out that the greater the number of filled elements, the higher the quality for the metadata record.

The other kind of tendency is related to semantic issues on the metadata elements. It is worth mentioning the considerations that the experts made on important free text elements such as the title and the abstract. Some experts stated that every title should answer, at least, the questions where, when, what and whom about the data; and that the abstract should describe, in a slightly broader way, the information which appears on the title, though they also thought that other issues can also be summarised there. Controlled elements such as the subject were found important as well, since they contribute to sort out subsets of topic-related records. The use of standardised thesauri, as the tool for filling in the subject element, was suggested as better than controlled lists. The rest of the semantic criteria focus their attention on more general aspects such as the coherence between the element and the information which it contains, the avoidance of duplicated information, the avoidance of contradictory information, the importance of precise information, the importance of homogeneity in the information among the metadata record set, and in a similar sense, the need for entity naming uniformity throughout the metadata record set and, finally, natural language semantic issues such as ambiguity which the experts recommend to minimise.



Structural Criteria

ISO Core filled
Number of filled elements
Data purpose
Spatial reference system
Distribution Format
Data responsible's information
Metadata creator
Access and usage data constraints
Data lineage
Data quality report

Semantic Criteria

Considerations on free text
Consideration on controlled elements
Coherence
Duplication of information
Contradictory information
Precise information
Uniformity of entity naming
Homogeneous information
Natural language semantics issues

Fig. 1. Compilation of criteria for the assessment of geographic metadata quality

Nonetheless, some other interesting criteria taxonomies can be proposed:

– according to the information type contained in the elements, the criteria may be sorted out into spatial (if they are related to spatial element types), textual (if related to textual element types) or temporal (if they deal with temporal element types).
– assuming that geographic metadata records do not usually appear in an isolated way, but form geographic thematic catalogues whose topics are diverse, from environmental aspects, to geographic images and cartography maps, there may exist quality criteria related to individual quality aspects, global quality aspects and both of them. In fact, it seems obvious that the quality of the individual records affects the perception on the repository. For instance, let us consider a metadata record set in which a high percentage of the records does not present an important, desired characteristic (i.e. presenting an accurate title field, presenting a correct topic-keyword classification and so on). Although some records fulfil the requirements, the overall impression on the set is likely to be of bad quality, circumstance that is confirmed because wrong records appear more frequently. Consequently, quality criteria which measure individual quality aspects, global quality aspects and both of them have to be taken into account.

Additionally, when studying the initial classification of the criteria (structural criteria and semantic criteria) more carefully, it can be stated that the semantic criteria merely determine the constituents of metadata without any regard to the quantity of each ingredient: they consider qualitative aspects of the metadata. On the contrary, the structural criteria give evidence of aspects which involve the measurement of quantity or amount which can be computed automatically.

In each engineering discipline, counting and measuring play an important role, because when it is feasible to measure the things that are being studied

Table 1. Proposal of geographic metadata metrics

Metric ID	*Metric name*	*Metric description*
Met1	purpose	Data purpose filled in
Met2	coreFilledPercentage	Percentage of the ISO Core filled in
Met3	alternateTitle	Number of words in the alternate title
Met4	numberOfFilledElements	Number of filled in elements
Met5	dataAccessConstraints	Data access constraints filled in
Met6	distributionFormat	Distribution format filled in
Met7	referenceSystem	Spatial reference system filled in
Met8	abstract	Number of words in the abstract
Met9	dataUpdateFrequency	Data and update frequency of the data filled in
Met10	title	Number of words in the title
Met11	responsiblesData	Information about the data responsible filled in
Met12	quality	Information about the data quality report filled in
Met13	lineage	Information about the lineage of the data filled in
Met14	metadataCreator	Information about the metadata creator filled in

and to express them in numbers, something is known about them. In addition, an important element in proving theories is provided by experiments, without measuring, experiments would be useless as an aid to natural scientists and engineers. After these considerations on the significance of measurement, it should be noted that there are important difficulties when measuring geographic metadata quality and, what is more, the engineering good practice of observing, counting and measuring regarding geographic metadata quality has so far been neglected. Undoubtedly, those quantitative criteria compiled (the structural criteria from Fig. 1) represent a good starting point in order to obtain metrics for geographic metadata quality. In Table 1, a list of 14 metrics for assessing quantitative aspects of geographic metadata quality is proposed. Some of the proposed metrics merely determine whether certain elements appear on the records (i.e. purpose, dataAccessConstraints or quality), others count the number of words per element (i.e. title, alternateTitle or abstract) and others try to determine the percentage of elements in the ISO Core that are filled in.

4 Analysing Geographic Metadata Quality Criteria

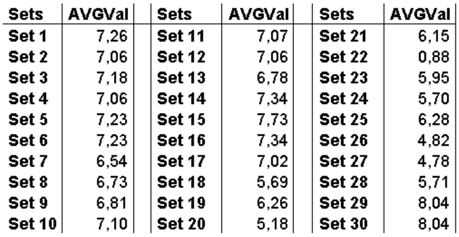
4.1 Methodology

With the aim of understanding the notion of geographic metadata quality, we decided to carry out another experiment which intended to discover the quantitative features which correlate significatively with good geographic metadata. Basically, the experiment consists of the following steps:

- select a sample of geographic metadata record sets
- ask the experts to assess the quality of the record sets with a numerical assessment

- compute the proposed metrics for the selected record sets
- analyse the correlation between the metrics and the assessments coming from the experts.

Table 2. The average value of the assessment per metadata record set



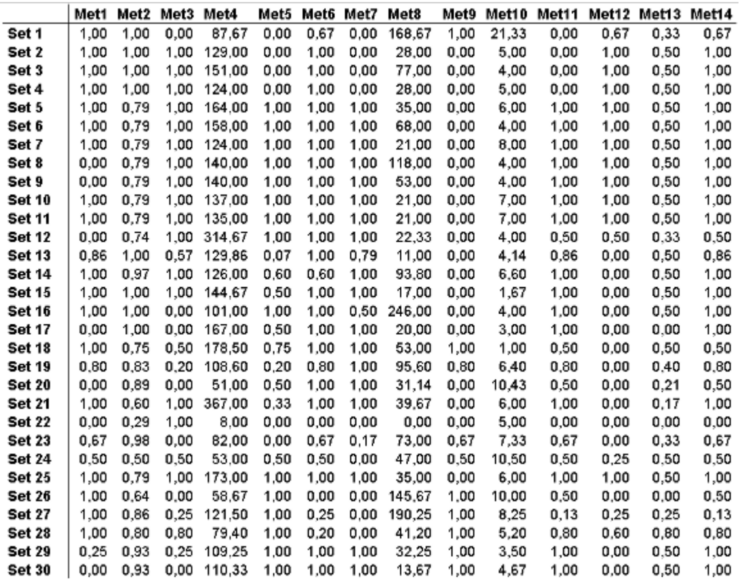
Sets	AVGVal	Sets	AVGVal	Sets	AVGVal
Set 1	7,26	Set 11	7,07	Set 21	6,15
Set 2	7,06	Set 12	7,06	Set 22	0,88
Set 3	7,18	Set 13	6,78	Set 23	5,95
Set 4	7,06	Set 14	7,34	Set 24	5,70
Set 5	7,23	Set 15	7,73	Set 25	6,28
Set 6	7,23	Set 16	7,34	Set 26	4,82
Set 7	6,54	Set 17	7,02	Set 27	4,78
Set 8	6,73	Set 18	5,69	Set 28	5,71
Set 9	6,81	Set 19	6,26	Set 29	8,04
Set 10	7,10	Set 20	5,18	Set 30	8,04

Because of the aforementioned reasons, the experiment was focused on the quality of the set rather than on individuals. Thus, 30 geographic metadata record sets of diverse cardinality were selected in order to carry out this experiment. They were compiled from different institutions: the Spanish National Geographical Institute, the French National Geographical Institute, several Spanish regional governments, some European institutions (such as the Joint Research Center) and the US Geological Survey. Their topics were Spanish, French and European cartography, Spanish and French hydrology, European LANDSAT images and orthoimages and geologic maps from the USA. The metadata record sets were all conforming to ISO 19115 with the exception of those from the US which were in CSDGM and were translated into the ISO 19115 standard by using the crosswalk described in [11]. Several experts from relevant public European organisations were asked to collaborate. Besides, the career backgrounds of the experts were rather heterogeneous: geographic, librarian and technologic.

The precise instructions given for the assessment were to assign a number from 1 (the lowest quality) to 10 (the highest quality) for each of the thirty metadata record sets and to write down an optional description for each of the assessments and a mandatory overall list of the assessment criteria. A form was given away in order to facilitate the noting down of those three elements. Two human-readable formats for the records were provided, one in HTML and another one in XML. A browser was recommended to visualise the records in the first case and the metadata edition tool CatMDEdit [12] in the second one. It is important to note, however, that neither evaluation criteria nor assessment recommendations were indicated to them. However, as geographic metadata represent the description of a particular geographic dataset and the dataset was not provided, the assessment was somehow constrained.

Once the results were compiled, the first necessary step for this statistical analysis was to obtain a unique assessment value per metadata record set. The assessments of the experts, however, differed slightly. The variation depended on the nature of the criteria chosen, since some of the experts were more concerned with structural aspects and others with semantic ones. An arithmetic average

Table 3. The numeric values of the metrics computed



	Met1	Met2	Met3	Met4	Met5	Met6	Met7	Met8	Met9	Met10	Met11	Met12	Met13	Met14
Set 1	1,00	1,00	0,00	87,67	0,00	0,67	0,00	168,67	1,00	21,33	0,00	0,67	0,33	0,67
Set 2	1,00	1,00	1,00	129,00	0,00	1,00	0,00	28,00	0,00	5,00	0,00	1,00	0,50	1,00
Set 3	1,00	1,00	1,00	151,00	0,00	1,00	0,00	77,00	0,00	4,00	0,00	1,00	0,50	1,00
Set 4	1,00	1,00	1,00	124,00	0,00	1,00	0,00	28,00	0,00	5,00	0,00	1,00	0,50	1,00
Set 5	1,00	0,79	1,00	164,00	1,00	1,00	1,00	35,00	0,00	6,00	1,00	1,00	0,50	1,00
Set 6	1,00	0,79	1,00	158,00	1,00	1,00	1,00	68,00	0,00	4,00	1,00	1,00	0,50	1,00
Set 7	1,00	0,79	1,00	124,00	1,00	1,00	1,00	21,00	0,00	8,00	1,00	1,00	0,50	1,00
Set 8	0,00	0,79	1,00	140,00	1,00	1,00	1,00	118,00	0,00	4,00	1,00	1,00	0,50	1,00
Set 9	0,00	0,79	1,00	140,00	1,00	1,00	1,00	53,00	0,00	4,00	1,00	1,00	0,50	1,00
Set 10	1,00	0,79	1,00	137,00	1,00	1,00	1,00	21,00	0,00	7,00	1,00	1,00	0,50	1,00
Set 11	1,00	0,79	1,00	135,00	1,00	1,00	1,00	21,00	0,00	7,00	1,00	1,00	0,50	1,00
Set 12	0,00	0,74	1,00	314,67	1,00	1,00	1,00	22,33	0,00	4,00	0,50	0,50	0,33	0,50
Set 13	0,86	1,00	0,57	129,86	0,07	1,00	0,79	11,00	0,00	4,14	0,86	0,00	0,50	0,86
Set 14	1,00	0,97	1,00	126,00	0,60	0,60	1,00	93,80	0,00	6,60	1,00	0,00	0,50	1,00
Set 15	1,00	1,00	1,00	144,67	0,50	1,00	1,00	17,00	0,00	1,67	1,00	0,00	0,50	1,00
Set 16	1,00	1,00	0,00	101,00	1,00	1,00	0,50	246,00	0,00	4,00	1,00	0,00	0,50	1,00
Set 17	0,00	1,00	0,00	167,00	0,50	1,00	1,00	20,00	0,00	3,00	1,00	0,00	0,00	1,00
Set 18	1,00	0,75	0,50	178,50	0,75	1,00	1,00	53,00	1,00	1,00	0,50	0,00	0,50	0,50
Set 19	0,80	0,83	0,20	108,60	0,20	0,80	1,00	95,60	0,80	6,40	0,80	0,00	0,40	0,80
Set 20	0,00	0,89	0,00	51,00	0,50	1,00	1,00	31,14	0,00	10,43	0,50	0,00	0,21	0,50
Set 21	1,00	0,60	1,00	367,00	0,33	1,00	1,00	39,67	0,00	6,00	1,00	0,00	0,17	1,00
Set 22	0,00	0,29	1,00	8,00	0,00	0,00	0,00	0,00	0,00	5,00	0,00	0,00	0,00	0,00
Set 23	0,67	0,98	0,00	82,00	0,00	0,67	0,17	73,00	0,67	7,33	0,67	0,00	0,33	0,67
Set 24	0,50	0,50	0,50	53,00	0,50	0,50	0,00	47,00	0,50	10,50	0,50	0,25	0,50	0,50
Set 25	1,00	0,79	1,00	173,00	1,00	1,00	1,00	35,00	0,00	6,00	1,00	1,00	0,50	1,00
Set 26	1,00	0,64	0,00	58,67	1,00	0,00	0,00	145,67	1,00	10,00	0,50	0,00	0,00	0,50
Set 27	1,00	0,86	0,25	121,50	1,00	0,25	0,00	190,25	1,00	8,25	0,13	0,25	0,25	0,13
Set 28	1,00	0,80	0,80	79,40	1,00	0,20	0,00	41,20	1,00	5,20	0,80	0,60	0,80	0,80
Set 29	0,25	0,93	0,25	109,25	1,00	1,00	1,00	32,25	1,00	3,50	1,00	0,00	0,50	1,00
Set 30	0,00	0,93	0,00	110,33	1,00	1,00	1,00	13,67	1,00	4,67	1,00	0,00	0,50	1,00

on the assessments was calculated in order to have a unique number per record set (see Table 2, note that again the values range from 1, the lowest quality, to 10, the highest quality).

The 14 metrics were computed for each of the 30 sets. The process consisted in computing the metrics for each of the records and then computing the average of those values to obtain the metric for the record set (see Table 3).

One way of studying the correlation of the metrics and the metadata record sets quality might be by determining the main source of variation in the metrics. This study was carried out by developing a Principal Component Analysis (PCA) [13]. The PCA is a mathematical procedure that transforms a number of variables into a smaller number of uncorrelated variables known as principal components (PCs). The first principal component (PC1) accounts for as much of the variability in the information as possible, and each succeeding component accounts for as much of the remaining variability as possible. The aim of this procedure is to reduce the dimensionality of data and to identify new meaningful variables. The relationship between the metadata quality values, coming from the assessments of the experts, and the principal component scores, obtained from the metrics, were studied through correlation analysis.

4.2 Results

Only the first component extracted from the PCA, which explained 32.2% of the observed variance (eigenvalue = 4.5), was significantly correlated with the metadata quality values (assessments). This correlation was strong and negative (R =

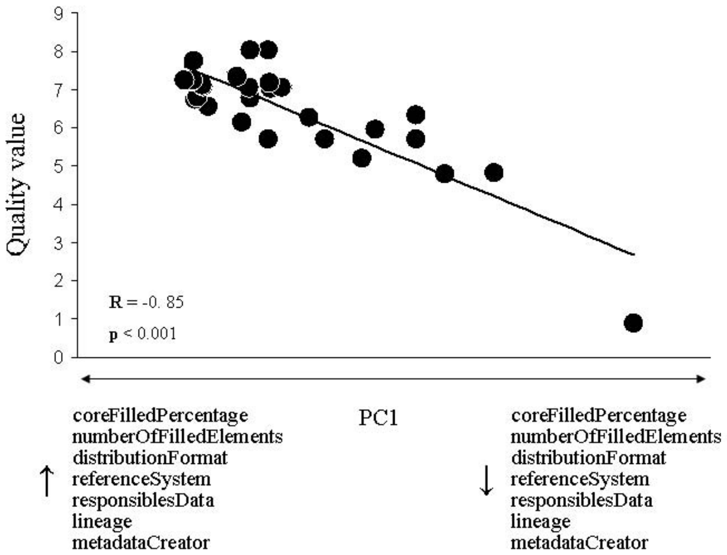


Fig. 2. The relationship between the quality values and the PC1

Table 4. The PCA *Factor loading*

Metric	PC1	PC2
purpose	-0.158	-0.603*
coreFilledPercentage	-0.486*	-0.595*
alternateTitle	-0.409	0.368
numberOfFilledElements	-0.794*	-0.221
dataAccessConstraints	-0.395	0.169
distributionFormat	-0.840*	0.015
referenceSystem	-0.710*	0.439*
abstract	-0.150	-0.781*
dataUpdateFrequency	0.441	-0.470*
title	0.424	-0.383
responsiblesData	-0.632*	0.269
quality	-0.402	-0.123
lineage	-0.616*	-0.305
metadataCreator	-0.853*	-0.193

-0.85) as Fig. 2 shows. The *factor loading* of the PCA reflects (Table 4) that this component (PC1) was significantly correlated with the metadata metrics: *coreFilledPercentage*, *numberOfFilledElements*, *distributionFormat*, *referenceSystem*, *responsablesData*, *lineage* and *cataloguersData*. The numerical values represent the correlation degree between the metrics and the PCs and the symbol * repre-

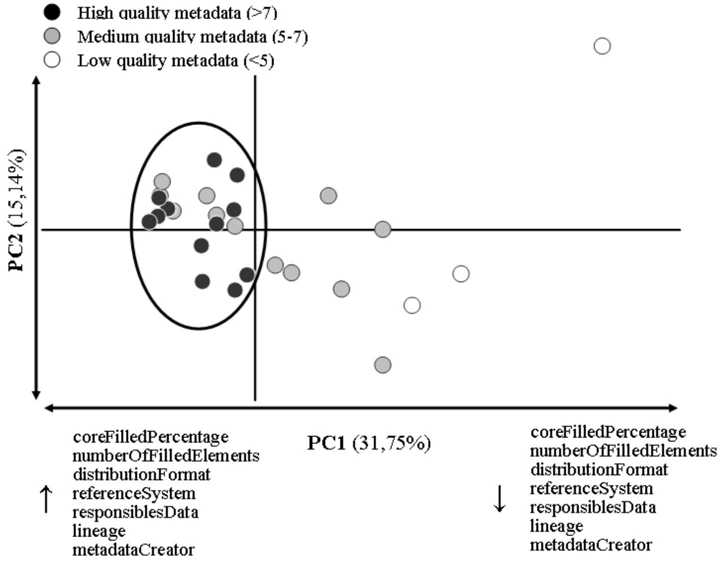


Fig. 3. Distribution of the different metadata record sets in relation to the PC1 and PC2

sents that there exists significant correlation (p <0.001). Thus, it can be concluded that these metadata metrics could be used as indicators of geographic metadata quality. If the value of the metrics increases, the quality of the record set increases as well. Nevertheless, the rest of the metrics were not significantly correlated and, consequently it cannot be statistically determined whether they have influence on the quality.

The first two components obtained through the PCA (PC1 and PC2) were used to represent the record sets in two dimensions (see Fig. 3). Metadata record sets were sorted into three groups according to the degree of their quality value degree (high quality, >7; medium quality, 5-7 and low quality, <5). The highest quality group appears associated to low values of PC1 and the lowest quality group with high values of this component.

According to Fig. 3, it is important to note that:

- high quality metadata record sets appear quite near among them and far way from poor quality metadata record sets
- high quality metadata record sets and some medium quality metadata record sets appear near what may suggest that the significantly correlated metrics do not determine quality completely and some other indicators such as those with semantic dimension take also an important role.

It can be stated that within this metadata set sample, the quality of the sets can be predicted by computing the correlated metrics. Thus, high values of the metrics involves medium-high quality and low values of them, low quality.

5 Conclusions

This work has presented early results from a series of experiments on identifying the quality of geographic metadata. The paper has proposed a quantitative method for quality assessment. The method is developed in two phases. Firstly, a list of geographic quality criteria was compiled from an opinion poll conducted to several experts of the area. The criteria were primarily classified into structural and semantic, though some other taxonomies were also described. The structural criteria give evidence of certain aspects which involve the measurement of quantity or amount which can be computed automatically. Derived from those criteria, a list of 14 geographic metadata metrics was proposed. Secondly, a statistical analysis was carried out on a selection of 30 geographic metadata record sets. The experiment, by developing a Principal Component analysis, studied the relationship between the 14 metrics, which were computed for each record set, and the assessments made by some experts. As a result, it was observed that some metrics could be used as indicators of geographic metadata quality and, within the selected 30 record sets, the geographic metadata quality could be predicted by computing those metrics: high values of the metrics involve medium-high quality and low values of them, low quality.

As further work and in order to validate these results and to generalise them, the experiments should be carried out with an extended metadata corpus. Additionally, it would be interesting to investigate whether metadata quality metrics can be applied to the development of more efficient information retrieval ranking algorithms. It is expected that quality metrics can play an important role in computing the relevance of the resource described.

References

1. Barton, J., Currier, S., Hey, J.: Building Quality Assurance into Metadata Creation: an Analysis based on the Learning Objects and e-Prints Communities of Practice. In: Proceedings of the 2003 Dublin Core Conference: Supporting Communities of Discourse and Practice-Metadata Research and Applications. (2003) ISBN 0-9745303-0-1.
2. Holden, C.: From Local Challenges to a Global Community: Learning Repositories and the Global Learning Repositories Summit. The Academic ADL Co-Lab (2003) Version 1.0.
3. Guy, M., Powell, A., Day, M.: Improving the Quality of Metadata in Eprint Archives. Ariadne Magazine (38) (2004) http://www.ariadne.ac.uk/.
4. Federal Geographic Data Committee (FGDC): Content Standard for Digital Geospatial Metadata, version 2.0. Document FGDC-STD-001-1998. Technical report (1998)
5. International Organization for Standardization (ISO): Geographic information - Metadata. ISO 19115:2003 (2003)
6. Armento, B., Terveen, L., Hill, W.: Predicting expert quality ratings of Web documents. In: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. Does "authority" mean quality?, Athens, Greece (2000) 296 – 303 ISBN 1-58113-226-.

7. Custard, M., Summer, T.: Using Machine Learning to Support Quality Judgements. D-Lib Magazine **11**(10) (2005) ISSN 1082-9873.
8. Hughes, B.: Metadata Quality Evaluation: Experience from the Open Language Archives Community. In: Proceedings of the 7th International Conference on Asian Digital Libraries (ICADL 2004). Number 3334, Lecture Notes on Computer Science. Springer-Verlag (2004) 320–329 ISBN 3-540-24030-6.
9. International Organization for Standardization (ISO): Information and documentation - The Dublin Core metadata element set. ISO 15836:2003 (2003)
10. Zhang, B., Gonçalves, M., Fox, E.: An OAI-Based Filtering Service for CITIDEL from NDLTD. In: Proceedings of the 6th International Conference on Asian Digital Libraries (IACDL 2003). Number 2911, Lecture Notes on Computer Science. Springer Verlag (2003) ISBN 3-540-20608-6, pp 590-601.
11. Nogueras-Iso, J., Zarazaga-Soria, F.J., Lacasta, J., Béjar, R., Muro-Medrano, P.R.: Metadata Standard Interoperability: Application in the Geographic Information Domain. Computers, Environment and Urban Systems **28**(6) (2004) 611–634
12. Nogueras-Iso, J., Zarazaga-Soria, F.J., Muro-Medrano, P.R.: Geographic Information Metadata for Spatial Data Infrastructures - Resources, Interoperability and Information Retrieval. Springer Verlag (2005) ISBN 3-540-24464-6.
13. Jolliffe, I.T.: Principal Component Analysis. 2nd edn. Springer Series in Statistics. Springer Verlag (2002)

Quality Control of Metadata: A Case with UNIMARC

Hugo Manguinhas and José Borbinha

INESC-ID – Instituto de Engenharia de Sistemas e Computadores,
Apartado 13069, 1000-029 Lisboa, Portugal
mangas@bn.pt, jlb@ist.utl.pt

Abstract. UNIMARC is a family of bibliographic metadata schemas with formats for descriptive information, classification, authorities and holdings. This paper describes the automation of quality control processes required in order to monitor and enforce quality of UNIMARC records. The results are accomplished by format schemas expressed in XML. This paper also describes the tools that take advantage of this technology to support the quality control processes, as also its actual applications in services at the National Library of Portugal.

1 Introduction

Descriptive metadata still plays a fundamental role in resource description services. Therefore, the quality of that metadata is also a key issue for the effectiveness of those services.

This paper addresses the problem of quality control of UNIMARC bibliographic records. This work was developed at the National Library of Portugal (BN), but the problem is addressed in a generic perspective, which means that the solutions here presented reused by any organisation dealing with the creation and processing of UNIMARC.

UNIMARC is a family of bibliographic metadata schemas with formats for descriptive information, classification, authorities and holdings. The UNIMARC adoption in Portugal followed the adoption in 1985 of UNIMARC as the international format for record exchange between national bibliographic agencies. Since then all BN cataloguing processes use UNIMARC as base format.

Until 2004, the validation and correction of records contained in the Portuguese union catalogue were performed by skilled professionals using cataloguing applications. Quality control processes are time consuming tasks that required a lot of attention, experience and expertise. BN requires quality control processes for two major divisions, record acquisition from cooperating libraries (libraries that signed a cooperation protocol with BN to maintain a single catalogue) into the Portuguese union catalogue (PORBASE) and another for union catalog maintenance.

Existing validation tools were embedded on legacy systems and couldn't be extended to accommodate format evolution. Most of them contained proprietary software systems that required continuous updates to current versions. Even software systems that enabled format update were unable to perform full UNIMARC compliance validation. On the other hand, we required validation tools, apart from our cataloguing systems, to satisfy a number of emerging quality control processes.

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 244 – 255, 2006.

To solve this problem, the National Library of Portugal (BN) started an activity to formalise a schema for the UNIMARC family of formats [1], ultimately supporting all the four formats (Bibliographic, Holdings, Authority and Classification) and their versions. In the following of that, a set of applications, named MANGAS (Manipulation and Management of Descriptive Metadata), were developed for the purpose of supporting automatic processes of quality control, using that schema.

This paper follows by introducing reader to the UNIMARC context and the problem of the quality control of UNIMARC records; it continues explaining all developed tools to aid theses processes and their applications over existing services on BN; finally, we describe the results already achieved and future work to be done.

2 UNIMARC

The primary purpose of UNIMARC is to facilitate the international exchange of bibliographic data in machine-readable form between national bibliographic agencies. UNIMARC belongs to a family of other MARC formats like MARC21 [2].

UNIMARC is intended to be a carrier format for exchange purposes. It does not stipulate the form, content, or record structure of the data within individual information systems. UNIMARC does provide recommendations only on the form and content of data for when it is to be exchanged.

Like other MARC formats, UNIMARC is a specific implementation of ISO 2709, an international standard that specifies the structure of records containing bibliographic data. It specifies that every bibliographic record prepared for exchange, conforming to the standard, must consist of a record label (a 24 character data element), followed by an undefined number of fields. A field is identified by a tag, a numeric three character code, and can be classified as a control or data field. Control fields contain a well defined set of character data, on the other hand, data fields may optionally contain one to two indicators (one alphanumeric character data, adding information about field content, relationships between the corresponding field and others, or about necessary data manipulation procedures) and subdivided into subfields. A subfield is identified by a one alphanumeric character symbol and can contain character data. Figure 1 shows the UNIMARC format class diagram.

The scope of UNIMARC is to specify the content designators (tags, indicators and subfield codes) to be assigned to bibliographic records in international machine-readable form and to specify the logical and physical format of the records. It covers monographs, serials, cartographic materials, music, sound recordings, graphics, projected and video materials, rare books and electronic resources.

The UNIMARC format was first published in 1977 under the title “UNIMARC - Universal MARC Format”. It was recommended by the IFLA Working Group on Content Designators set up by the IFLA Section on Cataloguing and the IFLA Section on Information Technology. It contained specifications for book and printed serial material and provisional fields for various non-book materials such as music, motion pictures, etc. Following editions added data fields required for cartographic materials, sound recordings, visual projections, video recordings, motion pictures, graphics, printed music, and microforms, and updated also several fields relating to serials and monographs.

Like most other formats, UNIMARC has evolved to accommodate new cataloguing practices related to existing or new bibliographic materials. These format changes involve defining additional fields, indicators, subfields and coded values where needed. These efforts, promoted by the Universal Control and International MARC Program, established an important normative support for the UNIMARC. In the 1985 IFLA Conference, UNIMARC was definitively adopted as the international format for record exchange between national bibliographic agencies and recommended as a model for further national MARC based formats in countries lacking an official format.

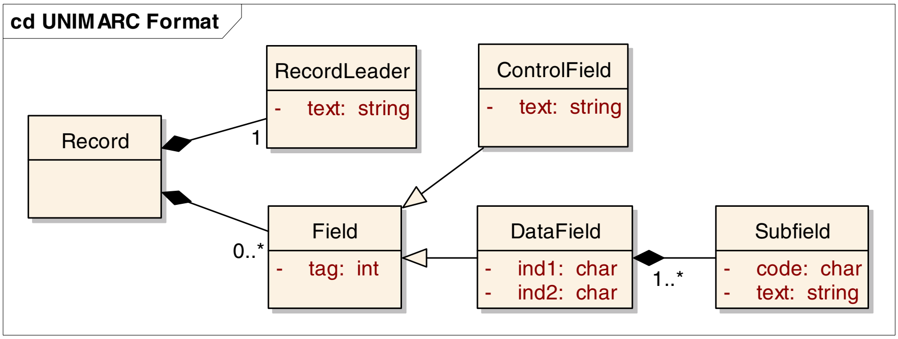


Fig. 1. UNIMARC Format Class Diagram

3 The UNIMARC Schema

UNIMARC is a format with a very complex structure. Besides the common syntactic rules for elements, attributes and values, it also defines semantic relations between them. These relations may even define the interpretation made for a given element or attribute. UNIMARC also requires grouping information in subsets of rules (aggregation of rules) that are required to represent blocks of fields. This requirement isn't essential for record validation but is important to define element semantic coupling.

Currently developed schema languages like XMLSchema [3], RELAX-NG [4] allow the definition of syntactic rules based on elements, attributes and values but lack semantic rules for defining relations between them. Schematron [5] is the only schema language that allows defining these semantic relations.

On the other hand, existing schema languages tend to evolve in time or be forgotten, and we need a stable format. We also required a schema language that was close to the concepts involved even if it required, for validation purposes, the conversion to another language (if possible).

We decided not to use any of the existing schema languages, but to develop our own schema language for the purpose of describing the UNIMARC formats, gathering all syntactic and semantic rules, each uniquely identified by an URN [6]. Figure 2 shows the corresponding class diagram for the UNIMARC format.

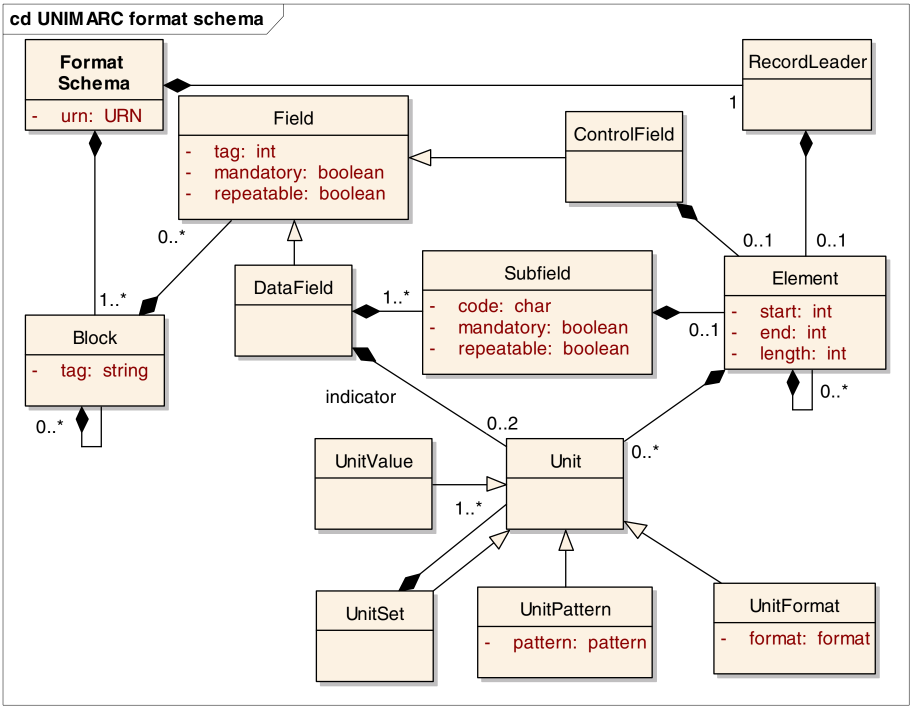


Fig. 2. Class Diagram for the Schema Language of the UNIMARC Format

Each UNIMARC format schema (Bibliographic, Authority, Holdings and Classification), and the respective versions, has its own format schema file with the corresponding format rules. Any format schema file can inherit the structural information of another UNIMARC format schema, making it possible to represent the format evolution by simply adding and replacing rules (overloading) in the newer versions.

4 Quality Control of UNIMARC Records

Quality control consists on a workflow of processes required for monitoring and enforcing quality over incoming and existing records. Figure 3 shows a possible quality control workflow containing validation, filtering, reporting and correction processes.

4.1 Validation

Validation is the first step and the basis for our quality control procedures. Like helping other quality control processes, other advantages emerge, like allowing end users to be instantly notified of cataloguing mistakes, their performance be measured (cataloguing procedures), common errors be detected and prevented, etc.

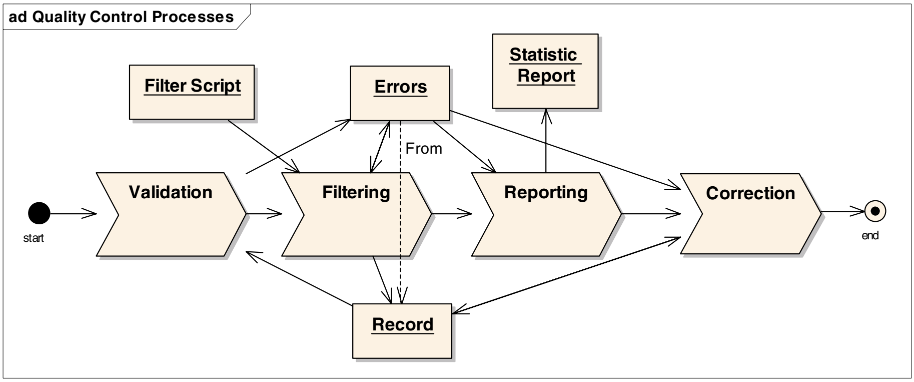


Fig. 3. Activity Diagram showing a possible quality control workflow

Schema validation is the process of checking vocabularies (well-formedness) and enforcing rules (constraints) embodied in schemas over metadata documents. The output of schema validation is a collection of identified errors (if any) in the source document. For our purpose we used schema validation to check UNIMARC records according to the UNIMARC format.

Among all the functional requirements common for schema validation we required a special feature, the generation of structured output errors that could enable more advanced quality control processes. These processes would use this error information to add more value. Among these are the generation of record error reports, record filtering based on record quality, or even record error correction.

At the moment, we are using, for performance reasons, our own schema validation tool. This enables us to validate a record without having to translate to a generic format schema whose constraints aren't optimized for record validation.

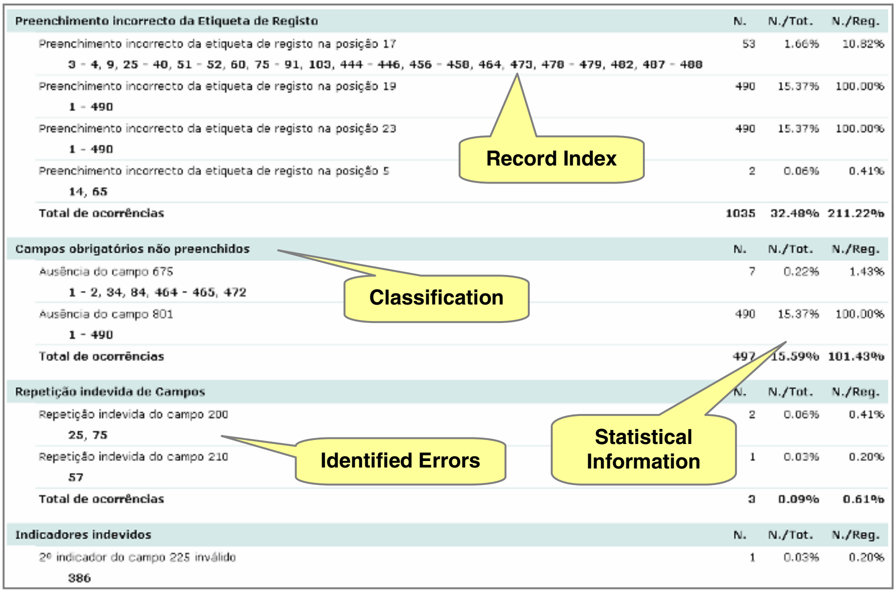
Nevertheless, any schema validation tool can be used for our purpose, as long as it supports error handling listener functionality for generation of structured output errors. On the other hand, any format schema can also be used, requiring it would be compliant with the UNIMARC format rules. A possible format schema would be Schematron based on the ability to define syntactic and semantic relations.

4.2 Filtering

Filtering consists on selecting records according to a given concern. On a quality control workflow it consists on distinguishing records according to the errors it contains and their characteristics.

Records can be divided according to the type of cataloguing skills required to perform a certain task or according to the level of concern it implies. Records requiring special attention of skilled professionals can be put apart from the remaining. Among these are records with complex errors, with errors affecting crucial information, etc. On the other hand, others with less complex errors can therefore be solved with help of common professionals and/or automatic procedures.

Filters can even select which records do not have the minimal required quality to be processed.



Preenchimento incorrecto da Etiqueta de Registo	N.	N./Tot.	N./Reg.
Preenchimento incorrecto da etiqueta de registo na posição 17	53	1.66%	10.82%
3 - 4, 9, 25 - 40, 51 - 52, 60, 75 - 91, 103, 444 - 446, 456 - 458, 464, 473, 478 - 479, 482, 487 - 488			
Preenchimento incorrecto da etiqueta de registo na posição 19	490	15.37%	100.00%
1 - 490			
Preenchimento incorrecto da etiqueta de registo na posição 23	490	15.37%	100.00%
1 - 490			
Preenchimento incorrecto da etiqueta de registo na posição 5	2	0.06%	0.41%
14, 65			
Total de ocorrências	1035	32.48%	211.22%
Campos obrigatórios não preenchidos	**N.**	**N./Tot.**	**N./Reg.**
Ausência do campo 675	7	0.22%	1.43%
1 - 2, 34, 84, 464 - 465, 472			
Ausência do campo 801	490	15.37%	100.00%
1 - 490			
Total de ocorrências	497	15.59%	101.43%
Repetição indevida de Campos	**N.**	**N./Tot.**	**N./Reg.**
Repetição indevida do campo 200	2	0.06%	0.41%
25, 75			
Repetição indevida do campo 210	1	0.03%	0.20%
57			
Total de ocorrências	3	0.09%	0.61%
Indicadores indevidos	**N.**	**N./Tot.**	**N./Reg.**
2º indicador do campo 225 inválido	1	0.03%	0.20%
386			

Fig. 4. Excerpt from a Record Error Index Report

4.3 Reporting

Reporting consists on gathering all available information and organizing (filtering, classifying and sorting) according to a given concern. In quality control processes the information is focused on record content and record quality information.

These reports are the source for more elaborate kind of statistic reports required for a number of other activities like monitoring, evaluating process accuracy, bookkeeping, or even, identifying problematic catalogue sources. This can also be used to help existing and future quality control processes. In a moment basis, these reports are important for managing the correction effort.

At the moment we build three kinds of reports, a detailed report showing all identified error occurrences in a single report, an index report merging error types and displaying the corresponding record indexes (see figure 4), and a summary report that merges all error types and shows the statistic information only. All these reports are built in machine or human readable format.

4.4 Correction

The correction is a very delicate process, requiring a high level of experience and trust in the developed system. To satisfy this requirement we decided to split the process into three distinct activities. This way we separate our concern into small problems each with increasing level of responsibility. This increases the level of control over the processes (enabling monitoring) and consequently the level of trust on the overall process. This process is cyclic and will finish when no correctable error is detected. Figure 5 shows the corresponding workflow for this process.

The process can be full automatic or require human interaction and approval. Each activity works as a module. The interface for input and output are well defined in order to enable the improvement of each activity without influencing the others. As we already mentioned there are three distinct and complementary activities:

- a **correction analysis activity**, responsible for selecting the possible corrections that fit the given scenario (record and corresponding errors);
- a **correction decision activity**, responsible for deciding which correction is more appropriate to the scenario, and a last activity;
- a **correction performing activity**, which applies the selected corrections to the record.

In most of the cases, where an error is well known and does not involve unpredictable information, this process goes smoothly.

A **correction analysis activity** is responsible for selecting a number of acceptable error corrections (hints), within a knowledge base, which fit the corresponding scenario. This knowledge base is maintained by professionals and enriched with new knowledge every time a new type of error is identified and the corresponding solution (if possible) is built. This activity only selects possible solutions. No reasoning is done over what is the right solution to be made. The activity receives as input a record and the identified errors and produces for outputs a correction script.

The correction script is composed of a set of correction hints for each identified error. There can be more than one possible correction (hint) for a given error. Correction hints are composed of actions that can be applied to a record in a given scenario in order to repair it. They are classified by a certainty degree that ensures a level of trust. These actions can be of two different types (add or remove) and are defined by a XPATH pointer to the data source element. Add actions are also followed by XSLT construct data elements.

The **correction decision activity** is responsible for choosing the best correction from a given scenario. This process can be done automatically with help of a decision making support systems (DSS) or manually with human intervention. The DSS could be as simple as choosing the first correction hint available with the highest certainty degree, or as complex as, inferring or reasoning based on pass experience (with help of artificial intelligence algorithms like neural networks, predicate logic and so on). At the moment we choose the simpler one.

It is not the responsibility of the correction decision activity to predict further errors (errors emerging from applying the correction) but only that, the ones already detected, are solved.

New human or machine based decisions, made in this activity, can be used to enrich the knowledge base used on the first activity.

Finally, the **correction performing activity**, receives the correction script, compiles the script into XSTL transformations and applies them to the record in order to repair it. After the record transformation, new unpredicted errors could occur emerging from relations between the changed information and existing information. To solve this problem the workflow must be cyclic and will end when no repairable error exists.

This workflow becomes more effective every time it is run. As new errors are detected, analysed and corrected, the process becomes more accurate.

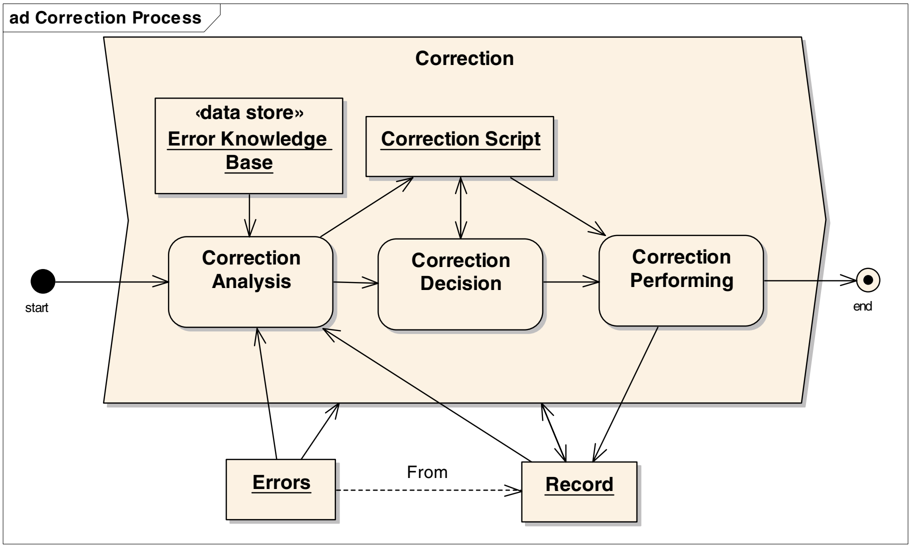


Fig. 5. Activity Diagram showing the correction process

5 Tools

A set of applications, named MANGAS, were developed to aid this quality control processes like validation, reporting, filtering and correction. These applications were developed essentially based on the end user profile. Common users, professionals or systems can use specific applications to perform their work. Nevertheless all these tools use the same core infrastructure to perform the available functionality.

5.1 MANGAS Diag

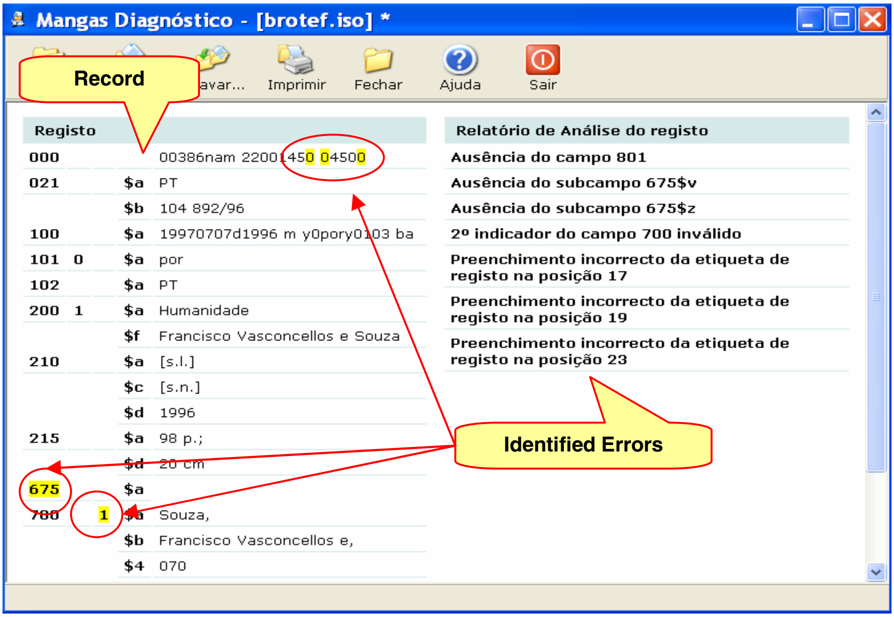
MANGAS Diag is a standalone tool developed to produce reports that gather the collected information. These reports can contain validation information or other record related information.

It is suited for common users who aren't familiar with the UNIMARC format and only require a way to build knowledge about their personal catalogues. This application is available as a standalone tool (see figure 6) or as a web service (located at http://diag.porbase.org).

5.2 MANGAS Workstation

The MANGAS Workstation is a tool for skilled professionals who are familiar with the UNIMARC format and require more complex functionality for their work. It is suited for professionals responsible for quality control procedures that require the ability to detect errors and act upon them.

Among the available functionality are the generation of custom reports (validation and content reports), different record views, record editing, record manipulation, record

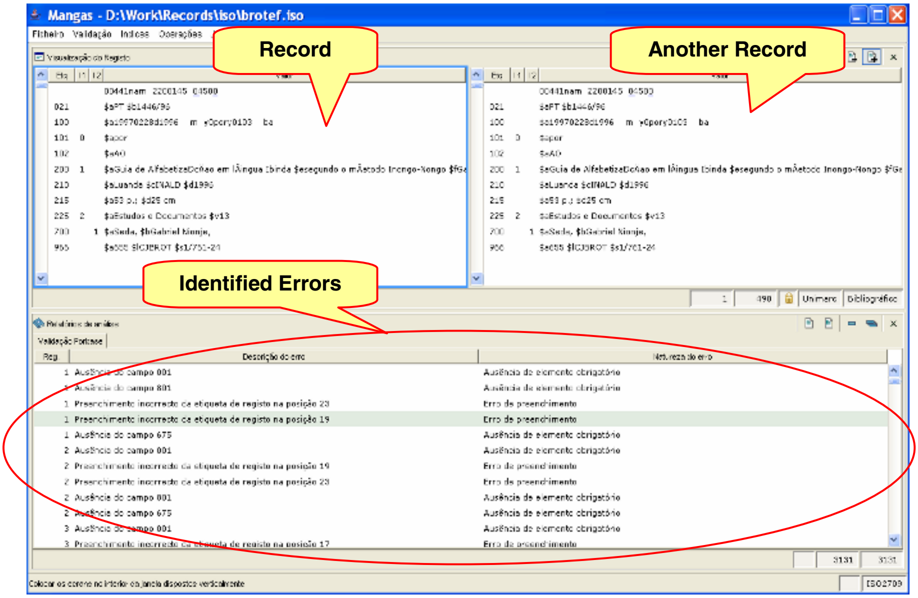


Registo			
000			00386nam 22001450 04500
021		$a	PT
		$b	104 892/96
100		$a	19970707d1996 m y0pory0103 ba
101	0	$a	por
102		$a	PT
200	1	$a	Humanidade
		$f	Francisco Vasconcellos e Souza
210		$a	[s.l.]
		$c	[s.n.]
		$d	1996
215		$a	98 p.;
		$d	20 cm
675		$a	
700	1	$a	Souza,
		$b	Francisco Vasconcellos e,
		$4	070

Relatório de Análise do registo
Ausência do campo 801
Ausência do subcampo 675$v
Ausência do subcampo 675$z
2º indicador do campo 700 inválido
Preenchimento incorrecto da etiqueta de registo na posição 17
Preenchimento incorrecto da etiqueta de registo na posição 19
Preenchimento incorrecto da etiqueta de registo na posição 23

Fig. 6. MANGAS Diag standalone tool

transformation, record error correction, search and replace functionality and printing capabilities. This application is only available as a standalone tool (see figure 7).



```
        00441nam  2200145  04500
021     $aPT $b1446/96
100     $a19970228d1996   m  y0pory0103   ba
101  0  $apor
102     $aAO
200  1  $aGuia de AlfabetizaÐção em lÃngua Ibinda $esegundo o mÃ©todo Inongo-Nongo $fGa
210     $aLuanda $cINALD $d1996
215     $a53 p.; $d25 cm
225  2  $aEstudos e Documentos $v13
700   1 $aSada, $bGabriel Ninnje,
955     $a055 $ICJBROT $s1/751-24
```

```
        00441nam  2200145  04500
021     $aPT $b1446/96
100     $a19970228d1996   m  y0pory0103   ba
101  0  $apor
102     $aAO
200  1  $aGuia de AlfabetizaÐção em lÃngua Ibinda $esegundo o mÃ©todo Inongo-Nongo $fGa
210     $aLuanda $cINALD $d1996
215     $a53 p.; $d25 cm
225  2  $aEstudos e Documentos $v13
700   1 $aSada, $bGabriel Ninnje,
955     $a055 $ICJBROT $s1/751-24
```

Reg.	Descrição do erro	Natureza do erro
1	Ausência do campo 001	Ausência de elemento obrigatório
1	Ausência do campo 801	Ausência de elemento obrigatório
1	Preenchimento incorrecto da etiqueta de registo na posição 23	Erro de preenchimento
1	Preenchimento incorrecto da etiqueta de registo na posição 19	Erro de preenchimento
1	Ausência do campo 675	Ausência de elemento obrigatório
2	Ausência do campo 001	Ausência de elemento obrigatório
2	Preenchimento incorrecto da etiqueta de registo na posição 19	Erro de preenchimento
2	Preenchimento incorrecto da etiqueta de registo na posição 23	Erro de preenchimento
2	Ausência do campo 801	Ausência de elemento obrigatório
2	Ausência do campo 675	Ausência de elemento obrigatório
3	Ausência do campo 001	Ausência de elemento obrigatório
3	Preenchimento incorrecto da etiqueta de registo na posição 17	Erro de preenchimento

Colocar os corone no interior da janela disposta verticalmente

Fig. 7. MANGAS Workstation stand alone tool

5.3 MANGAS Batch

The purpose of MANGAS Batch is to enable third party applications to use the MANGAS functionality to produce additional value. Currently this application is available as a standalone tool that can be called by a prompt in the local operating system, or be embedded as a library in a Java based application.

6 Services

For record acquisition, BN has created an elaborate workflow composed of quality control, duplicate entry checking and merger processes. This workflow is supported by a developed infrastructure named IRIS. This infrastructure works automatically but allows users to intervene whenever they decide or when their help is need. IRIS uses MANGAS tools for validation, filtering, reporting and correcting records (see figure 8).

In order to minimize BN acquisition effort, a number off services are provided to help partner libraries to fulfil this task. All MANGAS related tools used for quality control processes are available for download for local usage and a web service is available for prior validation (MANGAS Diag web service).

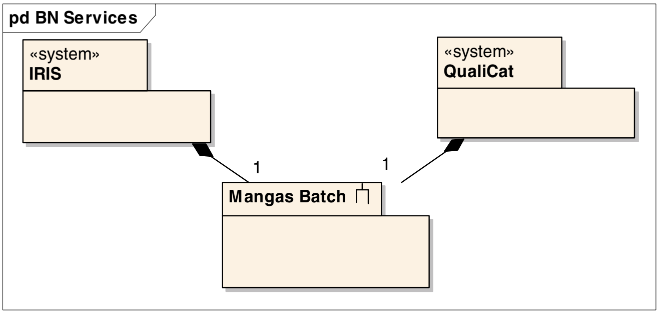
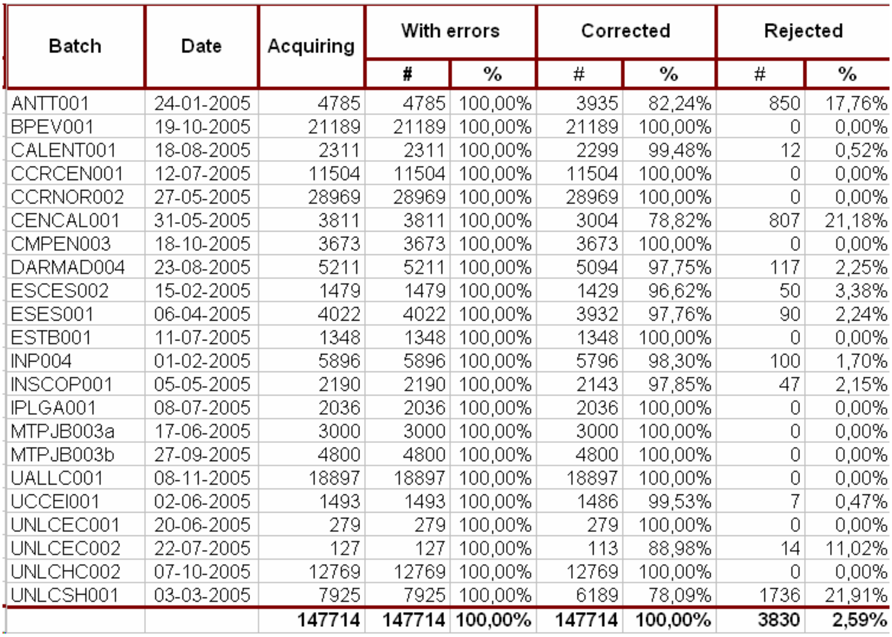


Fig. 8. BN Services System Model

For catalogue maintenance BN has developed a tool called QualiCat (web user interface only available on intranet) for the purpose of validating (using MangasBatch embedded in the system, see figure 8) all changed records during the day and reporting to the implied users in order to perform the necessary corrections. This tool runs every day at a scheduled time, usually after hours, to minimize interference with working processes.

7 Results

All this technology has been applied to BN quality control processes with excellent results. BN receives during the year a number of batches from other libraries to integrate with PORBASE they are submitted through IRIS (see figure 9). QualiCat runs every day



Batch	Date	Acquiring	With errors		Corrected		Rejected	
			#	%	#	%	#	%
ANTT001	24-01-2005	4785	4785	100,00%	3935	82,24%	850	17,76%
BPEV001	19-10-2005	21189	21189	100,00%	21189	100,00%	0	0,00%
CALENT001	18-08-2005	2311	2311	100,00%	2299	99,48%	12	0,52%
CCRCEN001	12-07-2005	11504	11504	100,00%	11504	100,00%	0	0,00%
CCRNOR002	27-05-2005	28969	28969	100,00%	28969	100,00%	0	0,00%
CENCAL001	31-05-2005	3811	3811	100,00%	3004	78,82%	807	21,18%
CMPEN003	18-10-2005	3673	3673	100,00%	3673	100,00%	0	0,00%
DARMAD004	23-08-2005	5211	5211	100,00%	5094	97,75%	117	2,25%
ESCES002	15-02-2005	1479	1479	100,00%	1429	96,62%	50	3,38%
ESES001	06-04-2005	4022	4022	100,00%	3932	97,76%	90	2,24%
ESTB001	11-07-2005	1348	1348	100,00%	1348	100,00%	0	0,00%
INP004	01-02-2005	5896	5896	100,00%	5796	98,30%	100	1,70%
INSCOP001	05-05-2005	2190	2190	100,00%	2143	97,85%	47	2,15%
IPLGA001	08-07-2005	2036	2036	100,00%	2036	100,00%	0	0,00%
MTPJB003a	17-06-2005	3000	3000	100,00%	3000	100,00%	0	0,00%
MTPJB003b	27-09-2005	4800	4800	100,00%	4800	100,00%	0	0,00%
UALLC001	08-11-2005	18897	18897	100,00%	18897	100,00%	0	0,00%
UCCEI001	02-06-2005	1493	1493	100,00%	1486	99,53%	7	0,47%
UNLCEC001	20-06-2005	279	279	100,00%	279	100,00%	0	0,00%
UNLCEC002	22-07-2005	127	127	100,00%	113	88,98%	14	11,02%
UNLCHC002	07-10-2005	12769	12769	100,00%	12769	100,00%	0	0,00%
UNLCSH001	03-03-2005	7925	7925	100,00%	6189	78,09%	1736	21,91%
		147714	**147714**	**100,00%**	**147714**	**100,00%**	**3830**	**2,59%**

Fig. 9. Report from 2005 record acquisition statistics

at a scheduled time, performing validation over all records that have been changed in PORBASE. PORBASE as an average of 492.58 changed records every day, 60.49% of those contain errors with an average of 1.67 errors per record, totalling 822.63 errors.

8 Work in Progress and Future Work

Concerning the quality control processes some aspects require more attention; for Validation purposes we need to develop converters from our schema language into existing generic schema languages (like Schematron) in order to enable third party systems, working with standard schemas, the usage of this technology; for Reporting, new customizable reports must be developed to satisfy different emerging information requirements; for Filtering we need to develop a way to customize filterers to apply in different situations; and finally, for Correction we need to improve the correction decision activity with a more advanced decision support system as well as register new types of errors and new solutions.

At a more practical level, we need to reevaluate the developed tools in order to identify functionality that needs to be improved and acquire new requirements for development of new functionality. Other services and systems concerning quality control processes must be reevaluated in order to take advantage of this technology.

References

1. IFLA – International Federation of Library Associations and Institutions (http://www.ifla.org).
2. LOC - MARC Standards, MARC in XML, September 2004 (http://www.loc.gov/marc/marcxml.html).
3. XML Schema (http://www.w3.org/XML/Schema).
4. RELAX-NG (http://relaxng.org).
5. Schematron (http://www.schematron.com).
6. Moats, R., "URN Syntax", RFC 2141, May 1997.

Large-Scale Impact of Digital Library Services: Findings from a Major Evaluation of SCRAN

Gobinda Chowdhury, David McMenemy, and Alan Poulter

Department of Computer and Information Sciences, University of Strathclyde,
Glasgow G1 1XH, UK
{gobinda.chowdhury, david.mcmenemy,
alan.poulter}@cis.strath.ac.uk

Abstract. This paper reports on an evaluation carried out on behalf of the Scottish Library and Information Council (SLIC) of a Scottish Executive initiative to fund a year's use of a major commercial digital library service called SCRAN throughout public libraries in Scotland. The methodology used for investigating value for money aspects, content and nature of the service, users and usage patterns, the effects of intermediaries (staff in public libraries), the training of those intermediaries and project rollout is given. Conclusions are presented about SCRAN usage and user and public library staff reactions.

1 Introduction

Even after a decade of intensive research and development activity, evaluation of large-scale digital library application and use still remains problematic. The ultimate goal of a digital library evaluation is to study how digital libraries are impacting on, and hopefully transforming, information seeking and use, research, education, learning and indeed the very lives of users. Several online bibliographies on digital library evaluation are now available (see for example, DELOS WP7 [1]; Neuhaus [2]; Giersch, Butcher and Reeves [3]; and Zhang [4]). Regular international workshops on digital library evaluation take place under the DELOS programme, and evaluation is a regular topic at all other digital library conferences. Several evaluation guidelines and methods have been proposed in course of evaluation projects like ADEPT [5], DELOS [6], eValued [7], JUBILEE [8], etc. Projects like eValued and HyLife [9] have developed toolkits and guidelines for evaluation of digital libraries. Many other researchers and institutions have also produced guidelines and toolkits for digital library evaluation. See for example: Reeves, Apedoe, and Woo [10]; Nicholson [11]; Borgman [12]; Blandford [13]; Blandford and Buchanan [14]; Blandford et al, [15]; Choudhury, Hobs and Lorie [16]; Chowdhury [17]; Borgman and Larsen [18]; Jeng [19] and Saracevic [20, 21, 22].

This paper reports on a recently completed large-scale evaluation of a major commercial digital library service called SCRAN (http://www.scran.ac.uk). This evaluation is unique for a number of reasons. First, it is an evaluation study of a large, nationwide, commercial digital library service, which was funded by the Scottish Executive to provide a specific range of services for all Scottish public libraries for one year, with the

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 256 – 266, 2006.

total cost of the project amounting to £123,900. Second, the outcome of the evaluation would determine whether Scottish Executive funding continued, thus it was necessary to ascertain the success or failure of the initial funding in value for money terms. Third, the evaluation was large-scale in that there are 557 public libraries in Scotland which attract over 31 million visits per annum. We would argue that the funding of access to a commercial digital library service by a national government for all citizens is hitherto a unique event in the development of very large-scale digital library services and needed to be evaluated extremely carefully, bearing in mind the complex social, economic and political aspects of the project. The evaluation however could not follow previously tried and tested well-trodden routes, for example by looking in detail at features like usability of individual pages in controlled conditions using a selected group of volunteers acting as users. It had to survey a large and diverse clientele of public library users to whom a large-scale digital library service was but one of a competing portfolio of services. Users could not be expected to recognize the uniqueness of the digital library service nor would its novelty alone give it any extra weight in their opinions. Public library staff, although being library professionals with an understanding of digital library services, would see it simply as yet another new service they had to support and deliver and would not give it any special treatment, apart from marketing it in the standard way as a new service. Finally, funders would not be looking for the meeting of research aims or achievement of good design but rather on visible take up and usage by the public vis a vis existing services and the reaction expressed by professional public library staff involved in its delivery.

A specific methodology was developed that addressed a number of issues including value for money aspects, content and nature of the service, users and usage patterns, the effects of intermediaries (staff in public libraries), the training of those intermediaries and project rollout. The paper briefly discusses the nature of the SCRAN service followed by the detailed methods used in the evaluation; major findings of the evaluation are then discussed with some critical comments that may be useful for the future design and management of large-scale digital libraries.

2 Background to SCRAN

SCRAN began in 1996. Its name came from an abbreviation of its initial purpose (Scottish Cultural Resources Access Network) but was also a reference to the Scottish word ‘scran’, which meant ‘food’ and ‘gather together’, very appropriate for a digital cultural portal. Resources were acquired through different stages of growth. The first batch came from Millennium funding in conjunction with the National Museums Service/National Library of Scotland. The actual digitisation of resources was outsourced. The second batch of resources came from NOF (National Opportunities Fund) funding for Resources for Learning in Scotland (http://www.rls.org.uk/). Other organisations provided resources, which SCRAN digitised and mounted and stored for fast access. SCRAN is essentially a federated database of resources from a variety of sources, some of which are commercial organizations, for example The Scotsman and Herald newspapers.

Over its history, SCRAN has accumulated a unique set of skills in digitisation and digital preservation. All of SCRAN's resources have copyright clearance for general use but with specific privileges for subscribers. SCRAN is currently working with the British Museum and the Scottish Motor Museum to acquire more resources.

Individual resource records are in Dublin Core format. Place names are provided by contributing institutions and can be variable as different institutions use different rules. SCRAN have tagged about 170,000 records in the past year with Ordnance Survey [the UK's national grid location system] co-ordinates. Geographic search allows linkages between areas and their sub-areas. There is no generic vocabulary or taxonomy for the vast range of subjects in SCRAN and contributing institutions themselves have no agreed system, which has the potential to influence the ability to efficiently search the resource. SCRAN are working with the Royal Commission on the Ancient and Historical Monuments (RCAHMS) and the National Museums of Scotland on a joint thesaurus for Scottish cultural institutions. SCRAN employ the UK Learning Object Metadata (LOM) with Pathfinder packs and they have a full hierarchy of curriculum terms for the English and Scottish curricula. SCRAN have three staff working full time on metadata – two checking, correcting and adding to records, and a data officer managing quality and carrying out global updates. SCRAN's three educational officers look after LOM information.

At the time of the evaluation, SCRAN offered an extensive range of materials consisting of over 1.3 million records, with over 300,000 multimedia resources, to schools, libraries and higher education institutions. Although SCRAN has created many 'Pathfinder' packs of resources by topic, SCRAN's interface has been extended over time to allow users to develop a range of resource applications for themselves by means of personalization or customisation. Such user-created information is stored on SCRAN's servers so it will work anywhere and not just on a local machine. 'My Stuff' offers a basic level of personalisation, like bookmarking. 'Albums' are more sophisticated, allowing user editing features (e.g. the addition of captions).

The Scottish Executive funding for access to SCRAN had several agreed objectives, viz.:

- To provide licensed access to SCRAN for all Scottish local authority libraries
- To provide user names and passwords to all participating libraries, and authentication system including IP authentication where required.
- To deliver a programme of training information professionals in developing their own use of the resources and in assembling learning objects
- To provide multi user rights to SCRAN 'Albums', CD-ROMs and resources to all libraries
- To provide 'Albums' functionality with captioning and local output to personal mini-website for use by public library staff to create their own 'Collections' for users
- To provide unrestricted 24/7 access, free at the point of use, to multimedia resources
- To handle IPR management of all resources.

Project management was provided by SCRAN, in conjunction with representatives from public libraries and from the Scottish Library and Information Council (SLIC), which is an independent advisory body to the Scottish Executive on library matters (http://www.slainte.org.uk/slic/).

3 Evaluation Objectives, Methods and Tools

The main objective of the evaluation of SCRAN was to assess the value for money of the year-long public library license. Outcomes could either be recommending continued access at the same (or higher or lower) cost or to devolve responsibility for funding to library authorities or to recommend an alternative to SCRAN.

In order to find answers to these questions the following multi-stage methodology was adopted involving the following tasks:

1. A detailed and critical study of the SCRAN website
2. Visits to SCRAN headquarters to interview key personnel and to study useful documents
3. Extensive analysis of web logs and other usage statistics supplied by SCRAN
4. A survey of selected public library staff to understand how the service is used by the end-users with the perceived benefits, level of difficulties, and various issues
5. A survey of end-users to understand the usage patterns and level of satisfaction
6. An analysis of the case study materials promoted by SCRAN as examples of best practice
7. Analysis of minutes from Steering group and Project Group and relevant documentation from SLIC.

Each stage of the methodology aimed to find specific information about SCRAN that would answer specific questions relating to the evaluation of the service:

1. How much was SCRAN used? What factors affected usage?
2. What did users think of SCRAN?
3. What did public library staff think of SCRAN?

3.1 Factors Affecting Usage of SCRAN

In theory virtually anyone can be a SCRAN user – school children doing homework, students at all levels, community groups in public libraries and any individual. SCRAN has local resources for everywhere in Scotland; and these resources can have personal resonance for individuals, a service SCRAN label quite succinctly as 'reminiscence'. Originally SCRAN was a unique service, with no competitors. However this is no longer the case. There are a plethora of alternative channels for obtaining information that is available through SCRAN. For example public library services maintain local gateways giving alternative free access to Scottish digital resources. The Resources for Learning for Scotland Project (RLS) used the UK's New Opportunities Fund (NOF) funding to draw together contributors from across the public sector with the intention of the digital assets being freely available. Material held on RLS is a combination of SCRAN and RLS data, but while text-based information can be accessed freely, access to the full image requires SCRAN subscription. Other Scottish projects such as Am-Baile (http://www.ambaile.org.uk/en/highlights.jsp), Springburn Museum (http://gdl.cdlr.strath.ac.uk/springburn/), and Virtual Mitchell (http://www.mitchelllibrary.org/vm/) provide full access to all images and not just thumbnails. For general educational resources not related to Scotland, websites like the BBC's Learning Homepage (http://www.bbc.co.uk/learning/) provide stiff competition.

Transaction log data maintained by SCRAN for the months of January to May 2005 was made available. Over the five month period, the average number of sessions (defined as at least one access in a half-hour period) per branch on SCRAN for all Scottish public library authorities was 15. This equates to an average of 3 sessions per month for each branch in Scotland over the period. There were occasional peaks but these were found to correspond with periods of staff training on SCRAN. Thus SCRAN usage generally was very low.

One of the main objectives of the project funding was 24/7 access to SCRAN. However, the nature of library opening hours varies considerably across Scotland, meaning that 24/7 access may in fact equate to only a handful of hours of access per day for many members of the public. This should have been raised when negotiations on the funding of SCRAN were taking place and should have been a consideration from the point of view of pricing.

Low in-branch usage could potentially have been offset by high at-home usage. The ATHENS Access Management system (http://www.athens.ac.uk/) was SCRAN's preferred access model, whereby unique IP addresses were recognised and tied to authorised users. Because of the licensing requirements on SCRAN from contributors, each user must be identifiable so that should a resource be discovered being used illegally, SCRAN can tell the user to desist. A number of SCRAN's commercial and non-commercial contributors regularly trawl Google to see if their resources are being used illegally and let SCRAN know of any illegal uses they find. Whilst this is important for contributing commercial organisations like The Scotsman and Herald newspapers, it is not that important for public sector bodies like museums who are trying to increase access to their digital content.

However the implementation cost for this type of authentication approach outside of academia had made it prohibitive for local authorities to implement. Remote access to SCRAN (i.e. by a public library user from home) would be possible with a different type of authentication system. As an example, access for public library members to other databases such as NewsUK and Encyclopaedia Britannica has been set up, allowing library card holders to access the databases 24/7 from their home computers using only their library card number. This is true universal access and allows members of the public to access library services even when the building is closed.

Even within public libraries, the differing usage of IP addresses in different library authorities posed problems for accessing SCRAN, as while some used fixed IPs, some did not use them at all (North Ayrshire, Argyle and Bute plus parts of Highlands, are examples). A subsequent problem was that several IT departments within councils changed the IP addresses of the computers in their authority, causing authentication issues beyond the control of SCRAN. Access, then, in public libraries was by mainly menu and password authentication. Choosing the default authority level rather than a particular public library would hide access from that library and served to obfuscate usage logging.

The original focus of SCRAN was and continues to be aimed at schools, and there is certainly an argument for suggesting that its interface displays an age profile bias towards children. Some of the terminology used could be confusing to adults who have not undertaken training, and there may be issues for the casual adult browser who is drawn to the service via marketing material only to be faced with terminology such as: “Homework”, “My Stuff’, “Lucky Dip”, “Monkeying Around”, “Fun and Games” and “Sticky Pics”.

Each of these features in its own right is creative and greatly enhances the user experience of the site. However their use in a database aimed at a wider market than schools does need to be rethought. A more intuitive homepage for public libraries could have been developed, aimed at the wider range of ages and interests that this client market represents. Certainly, doubts about SCRAN’s interface were born out through the user questionnaire: 41% of users had difficulty in finding material on SCRAN using the simple search.

3.2 Public Library User Perceptions of SCRAN

A questionnaire survey was conducted with the users of SCRAN services in public libraries throughout Scotland. The main objective of the user survey was to ascertain public library users’ views on the service, problems encountered, and the users’ overall reactions to the service. A total of 351 responses to the user survey were received. The public library user survey indicated that 51% of respondents had never used the SCRAN service. This was not because of a lack of interest in computer-based services as such: 71% of respondents said they would use online services and only 8% said they would not. The remainder would use them but would prefer printed materials. There was no obvious bias against online services by facets like age or gender. Those who used the service were interested in many types of material available via SCRAN: materials that are unique to their locality, their country, or their family were the most popular choices.

Awareness of the SCRAN service within the library was high, despite less than 50% of respondents had actually used it. Comments received on using SCRAN included the following:

- “I find retrieval of results most problematic on SCRAN, there seems to be no consistency in what terms, names or subjects are used for indexing and retrieval”
- “In the past I have noted inaccuracies of information stored”
- “Sometimes filtering of results could be better. I tend to get lots of irrelevant material along with my search results”
- “I used SCRAN for the first time today and found it very easy to use and full of interesting information”

Some of the comments from users suggest that retrieval of results is an issue for many, and this reinforces the need for a richer metadata scheme.

In order to gauge value for money and willingness to pay, a question was asked that requested users to give a cost per session they would be willing to pay to access a service providing the types of material available on SCRAN. Over 58% of respondents indicated they felt such a service should be free, with a further 15% not wishing to put a figure on it. This suggests that public libraries would struggle if they wished to recoup from their users some of the outlay of a SCRAN subscription.

3.3 Public Library Staff Perceptions of SCRAN

Another feature that the usage log revealed was a discrepancy between different library authorities. A few (Fife, Borders, Aberdeen) appeared to be heavier users than the other authorities. It was felt that these differences in usage patterns among the various authorities may have been caused by several factors including effectiveness of staff training and staff attitudes towards new digital library services in general and SCRAN in particular, making some staff more committed to using SCRAN. The web-based questionnaire survey was designed to find out answers to these issues.

The survey was conducted via the Internet; a total of 419 responses were received. Interestingly, a high proportion of responses came from the 'committed' group of library authorities. The responses on initial training were very positive. It was noted that most popular internal method of marketing was word of mouth, making cascading of training to as many staff as possible an absolutely crucial issue for success. A variety of user marketing methods were noted, but none seemed to be predominant.

A number of respondents mentioned that in their experience an aging population might not be computer literate but showed a liking for reminiscence services. There was however a general awareness that SCRAN usage was very low, and lower in some authorities than others. Fife was known to be a high user but then as commented by the respondents "Fife always was keen on online services".

Finally, respondents were asked to indicate how much of an effect losing access to SCRAN would have on the library service. While being broadly warmly receptive to SCRAN the opinion of the largest group (37%) of respondents was that the effect of losing SCRAN would be limited, although a high percentage of respondents felt that the effect would be reasonable (29%), with a smaller number thinking the effect would be significant (21%).

Richer information about the staff attitudes towards the service, problems encountered while using the service on behalf of the users, etc., was ascertained though a series of interviews among library staff. User interviews were undertaken with a range of authorities, both from the group identified by usage statistics, and staff survey responses, as 'committed' users and those not in this group. The intention was to try to elucidate how staff viewed the effectiveness of training, the utility of new services delivered and value for money of the project. Altogether 17 individuals from five authorities were interviewed. Most were experienced library staff, with lengths of services ranging from 15 years up to 40; 11 were in professional grade, 6 para-professional. Their areas of responsibility ranged from managing one or more libraries, to managing a specific facet of service (e.g. ICT, specifically People's Network services, children's services, or local history) or being in customer-facing roles. All the staff had received an initial round of training and then a second round focusing on hands-on use and creating applications. All used links on local portals to promote SCRAN. A general issue was that local computer technical support was often over-stretched. One group commented that just getting bookmarks changed and icons placed on screens was extremely difficult as rights to do these tasks were maintained centrally.

All had engaged with ECDL (European Computer Drivers Licence, http://www.ecdl.co.uk/) and felt that they had the requisite IT skills for the job; although they recognised that they were continually being stretched. They also admitted to being stretched generally, because of shrinking staff numbers and an unchanging set of core tasks which were being added to by new tasks – "Staff are being hit by new initiative after new initiative, with no time to bed one down before the next arrives." However all interviewees appeared well motivated and keen to do the best they could for their users. All the respondents were engaged in making provision of local digitised services, in the areas of Scottish history, local history and family history. All agreed that genealogy and reminiscence especially were popular services. Most were using local portals to point to web resources or locally-mounted CD-ROMs.

Digitisation for local history collections was being attempted by some but costs and other difficulties meant that it was sometimes easier to ask users to go to a central library to consult originals. The drawback to this approach, as mentioned by some professionals, was that "some materials would sit in vaults forever". It was remarked that some popular sites (e.g. Statistical Accounts of Scotland online; http://stat-acc-scot.edina.ac.uk/stat-acc-scot/stat-acc-scot.asp) were moving to 'for pay' access which meant that users could not be directed to them anymore. SAS online still is a free service. It is the value added elements which are moving to a subscription service.

One issue with promotion that was raised suggested that SCRAN's name gave no indication of what it was. Also its name was easy to confuse with those of other services e.g. SCAN, the Scottish Archives Network. No one reported problems in using SCRAN and most praised the suite of tools which enabled customisation to be done. Most interviewees made only light use of SCRAN. The biggest driver of usage was SCRAN's newsletters which prompted a check of SCRAN for new features or materials. Some staff wanted access to SCRAN from home as there they would have had time to explore.

The interviews of staff indicated that they felt stretched, and while being appreciative of the SCRAN service were often not in a position to promote it. One interviewee also stated that she felt the service was only now beginning to be used by more staff as they were finding time to pass on the skills. A selection of the comments received from staff are summarised below:

- "Easy access and detailed information make this an invaluable tool for public use"
- "Excellent service that will grow in usefulness"
- "Money could have been better spent on subscriptions of our choice"
- "If SCRAN is allowed more time to develop (i.e. amass more material) its resources, it will become an increasingly useful tool for public library online services"
- "Not many people have used it. I think that it is a good site but with so many other sites on the Internet it is easy to find the images you're looking for elsewhere"
- "I think advertising of this tool is woefully inadequate, and it's not available on enough of our PCs"

4 Conclusion

When SCRAN began, it had a clear focus as an online archive of Scottish cultural materials. Now SCRAN offers a much wider range of services, and is downplaying its Scottish focus. Rather than being the sole provider in a focused market, SCRAN is trying to push into other markets. While SCRAN's major strength as a service is still in its Scottishness and its collection of Scottish material, by not concentrating on this SCRAN did not impact on the public in Scotland as a strong brand associated with Scottish culture. For marketing purposes in Scottish public libraries it would seem better to have used SCRAN's old full title, Scottish Cultural Resource Network, rather than the more gnomic 'SCRAN'. Marketing could have concentrated on this message; posters and rolling screen saver demos showing SCRAN resources for a locality, tailored for each public library in that locality, would have much more effectively revealed the depth of SCRAN's Scottish resource base. Behind the marketing should have been a range of new services that would engage users (for example picture 'tours' of a locality as it looked in the past, opportunities for individuals to contribute their personal resources to their public library, etc). Public libraries have been accused recently in the UK of not developing their image beyond being mere lenders of books, and the success of a new online service based around reminiscence would have been a great triumph. It is clear from comments quoted above that SCRAN has been the source of many moments of deep satisfaction for public library users and staff who found its material of local and personal relevance.

That there is value in SCRAN is fully supported by anecdotal evidence but that value is highly personal and transitory and not embedded as an expected feature of public library services. There was also a generally supported wish for a publicly funded archive of freely available digital resources commemorating and celebrating Scottish culture. This creates tension between SCRAN as a commercial entity and the publicly funded library service which supplies it with free content only to be charged later to access that same content. The irony is that SCRAN was formed with Millennium Commission funds initially, and has navigated into being a commercial subscription service, while maintaining some funding from public sources for specific projects from time to time, like the Scottish Executive funding making possible the initiative evaluated here. While there is nothing wrong per se in commercialising successful digital library projects, the commercial rationale ought not to conflict with the public interest, in this case for free public access to materials that are clearly owned by the public. The most negative comment made by public library staff was that “SCRAN is a product whose time has gone”. A counter example of the British Library’s website was cited as a free site which offered much the same facilities as SCRAN.

The issue of transferring ownership of a library’s own materials was of particular concern to public library staff. Without a SCRAN subscription, a library authority, and the public in the local communities it serves, could not view their own contributions to the SCRAN site. This means, in essence, that public library staff in that library authority would have to hand a list of the material they had provided, but members of the public served by that library authority would be blocked from accessing more than mere thumbnails of material that in theory belongs to them through their authority’s ownership of the material. This would happen in non-subscribing library authorities throughout Scotland. The ethos behind the Creative Commons (http://creativecommons.org/worldwide/scotland/) licensing based on Scottish law encourages the sharing of digital resources with the owner retaining IPR but allowing pre-agreed use of the resource. A distributed environment incorporating the Creative Commons license for Scotland would offer an opportunity to access digital material that was owned in the public domain.

There is a much bigger question of what that distributed environment would look like. What needs to be addressed is exactly how the Scottish digital heritage will be developed and accessed, whether that heritage should be held in a centralised commercial database or decentralised in a managed set of collections held by the public sector bodies that accumulate that heritage. We believe that provision of a national database of cultural materials could easily be provided by public bodies in Scotland if provided with appropriate funding. What is necessary is to ensure that rather than training for a specific service such as SCRAN, staff members in cultural institutions are trained to create and manage their own digital materials under a national umbrella. This would negate the need for the nation's cultural institutions to be reliant on commercial providers for delivering their digital materials, and instead allow the public to access their heritage free of charge.

Acknowledgments

The evaluation project of SCRAN was funded by the Scottish Library and Information Council. We are grateful to SLIC staff members, especially to Ms. Elaine Fulton, and SCRAN staff members for their cooperation and support in this project.

We would like to express our gratitude to all the users and LIS professionals who took part in the user survey. The full evaluation report is available at:

http://www.slainte.org.uk/Files/pdf/SLIC/scranevaluation.pdf

References

1. DELOS WP7 evaluation workpackage. Bibliography. 2005 http://dlib.ionio.gr/wp7/iterature.html
2. Neuhaus, C. "Digital library: evaluation and assessment bibliography". 2005. Available: http://www.uni.edu/neuhaus/digitalbibeval.html.
3. Giersch, S., Butcher, K. and Reeves, T. "Annotated bibliography of evaluating the educational impact of digital libraries", Online. 2003. Available: http://eduimpact.comm.nsdl.org/evalworkshop/eval_ann-bib_09-29-03.doc.
4. Zhang, Ying "Moving image collection evaluation: research background - digital Library evaluation". Available: http://www.scils.rutgers.edu/~miceval/research/DL_eval.htm.
5. Alexandria Digital Library Project. Research. 2005. Available: http://www.alexandria.ucsb.edu/research/eval/index.htm.
6. DELOS Network of Excellence on Digital Libraries. http://www.delos.info/
7. eValued. "An evaluation toolkit for e-library developments". Available: http://www.evalued.uce.ac.uk/
8. JUBILEE "JISC User Behaviour in Information Seeking: Longitudinal Evaluation of EIS" Available: http://online.northumbria.ac.uk/faculties/art/information_studies/imri/rarea/im/hfe/jub/hfjubilee.htm
9. The HyLife hybrid library toolkit. Available: http://hylife.unn.ac.uk/toolkit/
10. Reeves, T. C., Apedoe, X. and Woo, Y. Evaluating digital libraries: a user-friendly guide. NSDL.ORG. The University of Georgia, 2003
11. Nicholson, Scott "A conceptual framework for the holistic measurement and cumulative evaluation of library services", Journal of Documentation, Vol. 60 No. 2, 2004. pp.164 – 182.
12. Borgman, C. L. et al. "How geography professors select materials for classroom lectures: Implications for the design of digital libraries". In: Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries, Tucson, Arizona, USA. New York: ACM, 2004. pp.179-185.

13. Blandford, A. "Understanding user's experiences: evaluation of digital libraries". In: DELOS workshop on evaluation of digital libraries Padova, Italy. 2004. Available: http://www.delos.info/eventlist/wp7_ws_2004/Blandford.pdf
14. Blandford, A. and Buchanan, G. "Usability of digital libraries: A source of creative tensions with technical developments", TCDL Bulletin. 2003. Available: http://www.ieee-tcdl.org/Bulletin/current/blandford/blandford.html
15. Blandford, A., Keith, S., Connell, I. and Edwards, H. "Analytical usability evaluation for digital libraries: a case study". In: Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries. 2004. Available: http://portal.acm.org
16. Choudhury, S., Hobbs, B. and Lorie, M. "A framework for evaluating digital library services". D-Lib Magazine, Vol. 8 No. 7/8., 2002. Available: http://www.dlib.org/dlib/july02/choudhury/07choudhury.htm
17. Chowdhury, G.G. "Access and usability issues of scholarly electronic publications". In: Gorman, G.E. and Rowland, F. eds. Scholarly publishing in an electronic era. International yearbook of Library and Information management, 2004/2005. London: Facet Publishing, 2004. pp. 77-98.
18. Borgman, C.L.& Larsen, R. ECDL 2003 Workshop Report: Digital Library Evaluation - Metrics, Testbeds and Processes. D-Lib Magazine, 9(9), 2003. Available: http://www.dlib.org/dlib/september03/09inbrief.html#BORGMAN
19. Jeng, Judy "What is usability in the context of the digital library and how can it be measured? Information Technology and Libraries, Vol. 24(2), 2005. pp. 47-56.
20. Saracevic, T. "Digital library evaluation: Toward evolution of concepts -1- evaluation criteria for design and management of digital libraries", Library Trends. Assessing Digital Library Services, Vol. 49 No. 2, 2000. pp. 350- 369. Available: http://www.scils.rutgers.edu/~tefko/LibraryTrends2000.pdf
21. Saracevic, T. "Evaluation of digital libraries: an overview. Presented at the DELOS workshop on the evaluation of digital libraries". 2004. Available: http://dlib.ionio.gr/wp7/ws2004_Saracevic.pdf
22. Saracevic, T. "How were digital libraries evaluated?" In: Libraries in the Digital Age (LIDA 2005), 30May -3 June, Dubrovnik, Croatia. 2005. Available: http://www.scils.rutgers.edu/~tefko/DLevaluation_LIDA.pdf

A Logging Scheme for Comparative Digital Library Evaluation

Claus-Peter Klas[1], Hanne Albrechtsen[2], Norbert Fuhr[1], Preben Hansen[5], Sarantos Kapidakis[4], Laszlo Kovacs[3], Sascha Kriewel[1], Andras Micsik[3], Christos Papatheodorou,[4] Giannis Tsakonas[4], and Elin Jacob[6]

[1] University of Duisburg-Essen, Duisburg, Germany
[2] Institute of Knowledge Sharing, Copenhagen, Denmark
[3] MTA SZTAKI, Budapest, Hungary
[4] Ionian University, Kekryra, Greece
[5] Swedish Institute of Computer Science, Kista, Sweden
[6] Indiana University Bloomington, Bloomington, USA

Abstract. Evaluation of digital libraries assesses their effectiveness, quality and overall impact. To facilitate the comparison of different evaluations and to support the re-use of evaluation data, we are proposing a new logging schema. This schema will allow for logging and sharing of a wide array of data about users, systems and their interactions. We discuss the multi-level logging framework presented in [19] and describe how the community can add to and gain from using the framework. The main focus of this paper is the logging of events within digital libraries on a *generalised, conceptual level*, as well as the services based on it. These services will allow diverse digital libraries to store their log data in a common repository using a common format. In addition they provide means for analysis and comparison of search history data.

1 Introduction

Evaluation of digital libraries (DLs) aims at assessing their effectiveness, quality and overall impact. Analysis of transaction logs is one evaluation method that has provided DL stakeholders with substantial input for making managerial decisions and establishing priorities, as well as indicating the need for system enhancements. However, the quantitative nature of this method is often criticised for its inability to provide in-depth information about user interactions with the DL being evaluated. The results of logging studies are often localised and not easily interpretable outside the DL being investigated. The problem of generalisability is compounded by the absence of a standardised logging scheme that could map across the various logging formats being used. The development of such a scheme would facilitate comparisons across DL evaluation activities and provide the means for highlighting critical events in user behaviour and system performance.

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 267–278, 2006.

The logging scheme proposed in this paper is framed within the DL evaluation activities of the DELOS Network of Excellence. In our whitepaper on DL evaluation [9], we have identified the long-term goal of building a community for DL evaluation. Under the umbrella of an experimental framework that will serve as a theoretical and practical platform for the evaluation of DLs, the proposed logging scheme will allow for meaningful interpretation and comparison of DL transactions. By using this scheme, researchers will be able to extract re-usable data from the results of previous DL evaluations and to identify useful benchmarks, allowing for more efficient and effective design of evaluation studies.

In order to support the re-use of all possible evaluation data, we intend that the proposed scheme will account for all manner of data that can be collected from the user, the system and the user-system interaction. To this end, we are proposing a novel, multi-level logging framework that will provide complete coverage of the different aspects of DL usage. The main focus of this paper is the level of conceptual, generalised user actions, for which we describe the logging scheme in some detail. Based on this specification, we present tools that can help DL stakeholders to analyse the logging data according to their specific interests.

The remainder of this paper is structured as follows: A brief survey of related work on DL logging is given in Section 2. The levels of the proposed logging scheme are presented in Section 3, while Section 4 addresses the events that comprise the conceptual aspects of user actions. Section 5 presents an application of the logging scheme in the Daffodil DL. Using researchers as an example, we discuss how the various DL stakeholders can gain from a common logging scheme in Section 6. In conclusion, Section 7 summarises the arguments presented in this paper and outlines future work with logging schemes.

2 Related Work

In DL research, there is a tendency to identify patterns of information searching behaviour through the use of system features such as Boolean operators, use of ranking and sorting mechanisms, identification of the type and nature of queries submitted and analysis of users' distribution of these features [17, 24, 18].

Transaction logs also provide a useful resource for the remote evaluation of web-based information systems due to their ability to record every action that occurs during the user's interaction with a digital resource [6]. They are generally used to gather quantitative data and to optimise the generation of statistical indicators. Logs are widely employed in the DL domain because of the consistency they provide with respect to the conditions of data collection; but logging should be understood as only one aspect of the overall experimental framework, as suggested by [4]. For websites and for bibliographic systems such as OPACs, logs provide dependable and coherent information about the usage, traffic and performance of a given system. However, because logs can only partially reveal both the behaviour of the user and her level of satisfaction [7], richer information is frequently derived by applying other methods of analysis such as sequential pattern analysis and web surveys.

Goncalves and others proposed an XML log standard for digital library logging analysis [12, 11]. They concentrate on high level events that are generated by user actions, and describe events for searching, browsing, updating and storing actions. However, as pointed out by Cooper in [5], for a comprehensive transaction log analysis different sources of log data have to be taken into account. We distinguish between different aspects and levels of logging, as described in Section 3. Each of these levels is supported by a standard XML schema that defines the events of interest according to the special needs of the different stakeholders interested in the logging data.

In addition, we extend the original set of events that have been proposed to more comprehensively support the logging of actions allowed by modern DL services. For the analysis of logged events, we also differentiate between the various stakeholders of a DL system, including, for example, system owners, librarians, endusers, developers and researchers. Each stakeholder group requires a particular view on the logging data in order to address its specific information needs. To support these views on transaction log data and to combine and analyse data from the different aspects of logging, a number of tools is being developed.

There are other efforts to standardise aspects of transaction logs or to provide uniform classifications for their analysis. Yuan and Meadow [27] provide a codification of, among other aspects, the variables of participants in user studies.

3 Levels of Logging

When using transaction logs for evaluation, the main participants under survey are the user and the system, as well as the content that is being searched, read, manipulated, or created. The interaction between the system and the user can be examined and captured at various levels of abstraction.

1. System parameters
2. User-system interaction
   - UI events (keystrokes, mouse movement, etc.)
   - Interface action events (data entry, menu selections, etc.)
   - Service events (use of specific DL services)
   - Conceptual events (generic actions)
3. User behaviour

On a very low level, researchers or developers might be interested in getting information about key presses, mouse clicks and movements, and similar concrete interactive events. These can be grouped together into more meaningful interactive events with specific graphical tools or interface elements, several of which in turn might comprise an interaction with a digital library service like a citation search service. For all three of these levels of abstractions a common logging schema will help in comparison, e.g. of the usability of specific interfaces or services between different digital libraries.

An additional level of abstraction, that will be called the conceptual level, can be seen as an abstraction away from the concrete services of specific digital library systems towards generalised actions that users can take in using a DL system. Several generic services that could be captured at this level have been proposed in [12] and [11]. Herein we suggest a modified and expanded version of this selection.

By separating the log data into user data, system data and different abstraction levels of interaction – all connected by timestamps and identifiers linking a specific user with each event – the transaction logs can be used for following a users actions chronologically along a specific tier of the model. On the other hand, a comprehensive analysis of what exactly happens during a search action of the user across the various tiers is also possible.

Of course a digital library system might only capture data on a low level of abstraction to minimise the overhead incured by the logging, and use a mapping of UI events to more abstract events for later analysis.

3.1 System

The system tier of the logging model describes the changes of the system over time, as represented by various parameters of the hardware and software involved in providing the digital library services. These parameters can be captured directly by the backend, usually without involvement of the client software used to access the digital library.

Aspects of the system fall into two groups: static parameters like operating system, available memory, bandwidth or computing power, and dynamic parameters that change over time, usually in reaction to user action, like server load, amount of connected users, network traffic and ping times.

3.2 User Behaviour

At the other end of the spectrum is the tier representing the user and her behaviour, her changing and evolving cognitive model, and her interactions with the environment outside the digital library system. While system behaviour is usually easiest to capture, user behaviour can hardly be captured through transaction logs. Other methods are common within user studies, e.g. video capturing, think aloud studies, search diaries, interviews or questionnaires.

As with systems, aspects describing the user can be grouped into static and dynamic parameters. Static parameters like age, first language, professional or search background, social and organisational environment, usually don't change over the course of one session. Dynamic parameters on the other hand, might include frustration levels, or the user's progression along the stages of her information task.

3.3 User-System Interaction

The logging of the interaction of user and system in distributed systems is complicated by the fact that much of the interaction occurs at the client. Low-level interactive events therefore need to be captured at the site of the user and be transmitted to the server, which introduces difficulties for browser based user clients. In these cases only higher level interactions with corresponding events on the server side of the digital library might be logged and analysed later [14].

UI Level. On the lowest level of user-system or human computer interaction (HCI), events consist of single input events like keystrokes, mouse clicks or mouse movements. These are of interest for usage and usability studies of systems, and have been analysed and studied in HCI research for a long time [15]. This level corresponds to the keystroke level of the GOMS model [16] and can be combined with low-level events captured by other means (e.g. eye movements).

Interface Level. It is common practice in HCI research [13, 14] to group low-level events into higher-level abstractions for specifying more general models of the user-system interaction. A file selection that incorporates several mouse movements, clicks and textual input is combined into a single, abstract interface action. On this level e.g. the number and kind of interface actions necessary to complete a specific task can be compared between several digital library systems.

Service Level. In a further step of abstraction the service level combines several interface actions into more meaningful services provided by digital library systems. Most DL systems offer a number of different services that support searching and other tasks of the information seeking process, e.g. metadata or fulltext search, search for citations, annotating of documents, services for organising personal collections of DL objects, or for supporting the reviewing of documents. Depending on the tools and options offered by a specific system, it is possible that different combination of interface actions can be combined to use the same service.

Conceptual Level. While the first three levels of user-system interaction combine actions from the level below into a larger, higher-level action, the conceptual level represents an abstraction away from the concrete implementation of specific services, and tries to define generalised types of events. These generalised events or conceptual events represent the various actions that a user might pursue in a digital library system. While this list is probably not comprehensive, care has been taken to describe these actions in a general way that can be applied to most of todays DL systems.

4 Events on the Conceptual Level

On the conceptual level, we have identified several general event types that support comparative evaluation across DLs. These events are partially in line with the events proposed in [12, 11]. They identified some generic events – search, browse, update, store – and some higher concepts – annotate, filtering, recommending, rating, reviewing – which they call transactions. Our focus on the conceptual level represents the centrality of these events for log analysis and interpretation, because they indicate critical aspects of the user's interaction with the DL system and supply valuable data for rich interpretation of user behaviour. As has been highlighted in other DL logging studies [22], current approaches are often inadequate for capturing complex or abstract actions by the user and are therefore unable to elicit meaningful conclusions.

By logging data about general event types at the concept level, we provide a basis for *comparative evaluation* across DLs.

The event types and event properties that we have identified are neither fixed nor a comprehensive model of user-system interaction and should be viewed as recommendations that can also serve as discussion points. Each event consists of its own set of properties modelled in XML as sub-elements and attributes. Properties that are common for all events are:

- a unique session-id
- start-stop timestamps
- possible errors during the event
- a unique event-id
- a service name
- a cancelation indicator

In addition, each event as described herein also has an event-specific set of properties which are summarized in Table 1. If the collection of additional data about specific events is necessary for a study, it is suggested to extending standard event definitions through reference to an XML namespace that defines the new properties. For example, the standardised search event describes the search condition as a list of terms. However, many DLs allow for more complex query formulations such as Boolean queries, which could be stored in a specific field defined in the extended namespace as a sibling of the list-of-terms input element. Although comparison across DLs will only work on the list-of-terms property defined in the original namespace, researchers who require an extended view of logging events are guaranteed that no information will be lost in the process of comparison.

Search events represent any action of users that involve formulating a query or filter condition that is to be processed by a DL service against a collection. The collection can be the entire document space, already be pre-filtered or even be the result of a previous query. The system response consists of the subset (e.g. in the form of a ranked list) of objects from the initial collection that satisfy the given condition.

Navigate events represent actions that consist of selecting a specific item from a set of items or following a link to its target. This conceptual event includes the use of hyperlinks to navigate within a set of hypertext documents, but also navigating through a representation of a social network (e.g. of co-authors).

Inspect events capture the user actions of accessing a detailed view of a single object, like the metadata or fulltext of a result document. Similarly, looking up a definition of a term in an encyclopedia or dictionary, or semantic information from a thesaurus is seen as an *inspection* of this term.

Display events describe specific visualisations of DL objects. While the actual content of the presented information does not necessarily change, the change in view on this information is representated by a display event. This conceptual event encompasses actions ranging from a simple resorting of a list

Table 1. Subelements and properties of events

Search	query or filter condition, collection to be searched, system response
Navigate	link to be followed, current collection, system response
Inspect	object to be inspected, system response
Display	collection to be (re-) displayed, visual transformation or visualisation to be applied, sort criteria (optional)
Browse	collection to be browsed, method and dimensions of browsing, direction and distance that the view point is moved
Store	set of DL objects to be stored, target location, method of storage
Annotate	document (optionally part of) to be annotated, type and content of the annotation (may be one or more other DL objects)
Author	the new document, optionally identifier of changed document
Help	help request (optional), type and content of system suggestion
Communicate	type, content and recipients of message

of objects or the presentation of a set of terms in form of a tag cloud, up to complex visual representations of abstract information. In the interest of comparability, the use of a standardised taxonomy or classification of visualisation techniques is proposed, e.g. based on Shneiderman's task by data type taxonomy [25] or the classification of visualisation techniques in [3].

Browse events describe user actions that involve changing the view point on a set of DL objects without changing the visualisation or navigating to a different set of items. Typically these actions will involve scrolling in one or more dimension, using sliders to zoom in or out of a visualisation, or "thumbing" through a document. If the original set of documents has been split into several chunks, browsing might also describe moving from one chunk of the document set to the next or previous one.

Store events are actions of the user or the system that create a permanent or temporary copy of a digital library objects. This might be a digital copy to a clipboard for temporary storage during a search session or to a more permanent location either within the DL system or outside (e.g. on an optical medium or in a web storage). Storing a digital library object can also mean converting from digital to physical form (printing a document), or exporting to a special citation format.

Annotate events cover any user action that adds additional information to an existing DL object, which may be user-specific, shared among a group of collaborators, or visible system wide. The general annotation event includes marking entire documents or specific parts, adding ratings, tags [10] or textual comments like reviews or summaries to an object, or linking two or more DL objects.

Author events describe the creation of a new DL object or direct editing (not annotating) of an existing one. Authoring a document can include writing a review or another type of textual annotation, or can be part of creating a completely new document if supported by the digital library system.

Help events from the user's point of view can be of a passive nature, where the system provides unprompted suggestions to the user, or of a more interactive nature. In the latter case the user explicitely requests help, suggestions or recommendations about a specific or general topic, and the system generates a response to that. Unprompted help can take the form of recommendations about content, users or specific actions, or provide explanation about functions or system activity.

Communicate events capture events that occur during the communication between two or more users of the digital library system. The communication can be textual or include other media. This general event includes the direct sharing of a DL object with another user and sending messages to other users by using instant messaging, message boards or e-mail components of the system. More technically sophisticated systems might allow for sending voice or video messages between users as well. Digital library services for collaboration will typically also include means of communication for managing the collaboration.

Some of the events contain the actual digital library objects either in form of object identification like DOI, URL or URI. If such a unique identification is not possible, the main or all metadata fields should be provided to distinguish the objects.

5 Logging in Daffodil

Daffodil[1] [23, 21, 8] is a virtual DL targeted at strategic support of users during the information search process. For searching, exploring and managing DL objects Daffodil provides information seeking patterns that can be customised by the user for searching over a federation of heterogeneous digital libraries. Searching with Daffodil makes a broad range of information sources easily accessible and enables quick access to a rich information space.

5.1 Logging

Daffodil is an application consisting of a graphical user interface and back-end services written in Java. This makes logging either UI events within the client or back-end event triggered by user actions or by the system a much simpler process than for web-based DL systems. The logging service itself, depicted in Figure 1, is simplistic: the Daffodil user interface client and each Daffodil back-end service can send an event to the event logging service, which then stores the event.

Currently, Daffodil handles over 40 different events. The main groups of events are search, navigate and browse events; result events are generated by each

[1] `http://www.daffodil.de`

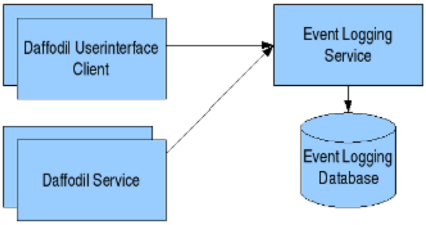


Fig. 1. Logging Service Model

of the system services (e.g., the thesaurus, the journal and conference browser, or the main search tool). The personal library supports store events as well as events involving annotation or authoring of objects.

Through the DAFFODIL service for log schema conversion, we are able to provide more than 100 MB of log data in the format of the proposed XML schema. The data will be anonymised and soon be made accessible for comparative analysis.

6 Analysis of Log Events

In order to analyse the logged events, we have assumed that different stakeholders need different views of the logging data; thus, a variety of analysis tools is required. We have identified the following DL stakeholders: *system owners*, *content providers*, *system administration*, *librarians*, *developers*, *scientific researchers*, and *end-users* of a DL.

A number of tools for facilitating analysis on log data in the new scheme habe already been implemented. The example statistics shown in Figures 2 and 3 were produced with the help of these tools.

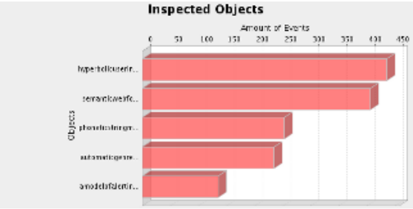


Fig. 2. Inspected Objects

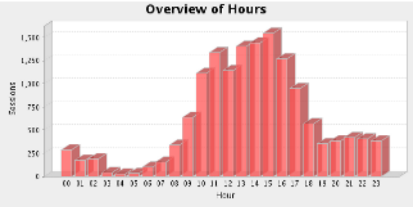


Fig. 3. Usage per hour

Having an experimental framework is a special boon for **researchers** in the field of digital libraries. With DAFFODIL and its use of the standardised logging scheme, a baseline system and a powerful toolbox for evaluation work can be provided. For research on higher concepts, the researchers do not have to develop or implement a complete setting, but can re-use the framework and build on it. At the system level, the efficiency of algorithms or the appropriateness of DL architectures can be evaluated and compared. HCI research on the usability of digital library interfaces as well can benefit from a framework that provides standardised, comparable logging data.

The major research focus of the Daffodil project has always been to place the user in the center of digital library and information retrieval research. Daffodil is based on the information search model by Marcia Bates [2, 1] which classifies search activities of users as moves, tactics, stratagems and strategies. The proposed logging scheme and the concept level is a natural extension of this original aim, as it provides a method to analyse the sequences of moves and tactics. As Wildemuth stated in [26]:

> While significant work has examined the individual moves that searchers make (Bates, 1979, 1987; Fidel, 1985), it is equally important to examine the sequences of moves made by searchers in order to understand the cognitive processes they use in formulating and reformulating their searches.

With regard to this goal, the logging scheme allows for studying the usage of the diverse services; furthermore, the whole search process/sequence can be analysed, from the initial formulation of a search query to its conclusion with the storage of DL objects for further use.

In addition to capturing and understanding users' search activities through analysis of logging data, recommendations can be made for re-use of search patterns. Kriewel [20] suggests that search paths discovered in analysis of logging data can be used as a basis for suggesting potential search steps to other users.

Of course the vertical look at the levels can also be under examination.

7 Summary and Outlook

In this paper, we have presented the first efforts to develop a standardised experimental framework for digital library evaluation. If the various DL stakeholders can form a community and agree upon models, dimensions, definitions and common sense criteria for the evaluation of DL systems, the process of evaluation will gain impetus among DL researchers. As an experimental platform for evaluation both within and across DL systems, application of the Daffodil framework can substantially advance this research area, since it currently provides functions and services that are based on a solid theoretical foundation and well known models. The proposed logging scheme is a first step intended to encourage evaluation of individual systems as well as comparisons across systems.

Most evaluation techniques require a great deal of preparation and effort and are thus not easily replicated. In case of online DL systems, this means that the results of an evaluation often reflect a past snapshot of the system. It is necessary to find ways for continuous, cost effective and (more or less) automated evaluation of digital libraries. The suggested logging scheme is a first step in this direction. Our group aims at establishing a community forum for evaluators in order to promote the propagation of various tools and approaches, and the exchange of experience.

As part of the effort to encourage a community forum for researchers interested in DL evaluation, we have published documentation for the logging scheme on the forum website[2]. The log analysis tools will follow soon, as they are under preparation. As a further step, large amounts of anonymised logging data from two DL services will be made available. Such primary data will help the community to improve the application of transaction logging and to compare and experiment with sample data. In the long term, we envision that this effort will evolve into a primary data repository, providing help for evaluators who want to find similar scenarios together with logging data. In other fields of research, such primary data repositories are already well established and play an important role in the conduct of research. (In fact, the provision of primary data frequently counts as a publication, creating further incentives for this kind of work.) An infrastructure of independent evaluator services, primary data repositories, logging tools and on-line questionnaires may provide computer-based support for some of the cost-effective and time-consuming tasks in evaluation and the community will gain sustainability. So we ask the community to form, discuss and agree upon a schema, to add data and service in order make a step forward in DL evaluation.

References

[1] M. J. Bates. Idea tactics. *JASIS*, 30(5):280–289, 1979.

[2] M. J. Bates. Information search tactics. *JASIS*, 30(4):205–214, 1979.

[3] E. Bertini, T. Catarci, S. Kimani, and L. D. Bello. Visualization in digital libraries. In *From Integrated Publication and Information Systems to Virtual Information and Knowledge Environments*, pages 183–196, 2005.

[4] J. Bertot, C. McClure, W. Moen, and J. Rubin. Web usage statistics: measurement issues and analytical techniques. *Goverment Information Quarterly*, 14(4):373–3951, 1997.

[5] M. D. Cooper. Design considerations in instrumenting and monitoring web-based information retrieval systems. *JASIS*, 49(10):903–919, 1998.

[6] M. D. Cooper. Usage patterns of a web-based library catalog. *JASIST*, 52(2):137–148, 2001.

[7] D. T. Covey. Usage and usability assessment: Library practices and concerns. Technical report, Digital Library Federation, 2002.

[8] N. Fuhr, C.-P. Klas, A. Schaefer, and P. Mutschke. Daffodil: An integrated desktop for supporting high-level search activities in federated digital libraries. In *Research and Advanced Technology for Digital Libraries. 6th European Conference, ECDL 2002*, pages 597–612. Springer, 2002.

[9] N. Fuhr, G. Tsakonas, T. Aalberg, M. Agosti, P. Hansen, S. Kapidakis, C.-P. Klas, L. Kovcs, M. Landoni, A. Micsik, C. Papatheodorou, C. Peters, and I. Solvberg. Evaluation of digital libraries, 2006.

[2] `http://www.is.informatik.uni-duisburg.de/wiki/index.php/JPA_2_-_WP7`

[10] S. A. Golder and B. A. Huberman. The structure of collaborative tagging systems. Technical report, Information Dynamics Labs, HP Labs, 2005.

[11] M. A. Goncalves, E. A. Fox, L. Cassel, A. Krowne, U. Ravindranathan, G. Panchanathan, and F. Jagodzinski. Standards, mark-up, and metadata: The xml log standard for digital libraries: analysis, evolution, and deployment. In *Proceedings of the third ACM/IEEE-CS joint conference on Digital libraries*, 2003.

[12] M. A. Goncalves, R. Shen, E. A. Fox, M. F. Ali, and M. Luo. An xml log standard and tool for digital library logging analysis. In *Agosti, Maristella (ed.) et al., Proc. of 6th ECDL*, pages 129–143, 2002.

[13] M. A. Hearst and M. Y. Ivory. The state of the art in automating usability evaluation of user interfaces. pages 470–516, 2001.

[14] J. Helms, D. Neale, and P. I. amd J.M. Carroll. Data logging: Higher-level capturing and multi-level abstracting of user activities. In *Proceedings of the 40th annual meeting of the Human Factors and Ergonomics Society*, 2000.

[15] D. M. Hilbert and D. F. Redmiles. Extracting usability information from user interface events. *ACM Computing Surveys*, 32(4):384–421, 2000.

[16] B. E. John and D. E. Kieras. Using goms for user interface design and evaluation: Which technique? acm trans. *ACM Trans. Comput.-Hum. Interact. 3*, 1996.

[17] S. Jones, S. J. Cunningham, R. McNab, and S. Boddie. A transaction log analysis of a digital library. *International Journal on Digital Libraries*, 3(2):152–169, 2000.

[18] H. Ke, R. Kwakkelaar, T. Tai, and L. Chen. Exploring behavior of e-journal users in science and technology: Transaction log analysis of elsevier's sciencedirect onsite in taiwan. *Library & Information Science Research*, 24(3):265–291, 2002.

[19] C.-P. Klas, H. Albrechtsen, N. Fuhr, P. Hansen, E. Jacob, S. Kapidakis, L. Kovacs, S. Kriewel, A. Micsik, C. Papatheodorou, and G. Tsakonas. An Experimental Framework for Comparative Digital Library Evaluation: The Logging Scheme. In *JCDL, Short Paper*, 2006.

[20] S. Kriewel. Finding and using strategies for search situations in digital libraries. Bulletin of the IEEE Technical Committee on Digital Libraries (to appear), 2005.

[21] S. Kriewel, C.-P. Klas, A. Schaefer, and N. Fuhr. Daffodil - strategic support for user-oriented access to heterogeneous digital libraries. *D-Lib Magazine*, 10(6), June 2004.

[22] B. Pan. Capturing users behavior in the national science digital library (nsdl). Technical report, NSDL, 2003.

[23] A. Schaefer, M. Jordan, C.-P. Klas, and N. Fuhr. Active support for query formulation in virtual digital libraries: A case study with DAFFODIL. In *Proc. of 7th ECDL*, 2005.

[24] M. Sfakakis and S. Kapidakis. *User Behavior Tendencies on Data Collections in a Digital Library*, volume 2458 of *Lecture Notes In Computer Science*, pages 550–559. Springer-Verlag, Berlin; Heidelberg, 2002.

[25] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. Technical Report CS-TR-3665, University of Maryland, Department of Computer Science, July 1996.

[26] B. M. Wildemuth. The effects of domain knowledge on search tactic formulation. *JASIST*, 55(3):246–258, 2004.

[27] W. Yuan and C. T. Meadow. A study of the use of variables in information retrieval user studies. *JASIS*, 50(2):140–150, 1999.

Evaluation of Relevance and Knowledge Augmentation in Discussion Search

Ingo Frommholz and Norbert Fuhr

University of Duisburg-Essen
Duisburg, Germany
ingo.frommholz@uni-due.de,
fuhr@uni-duisburg.de

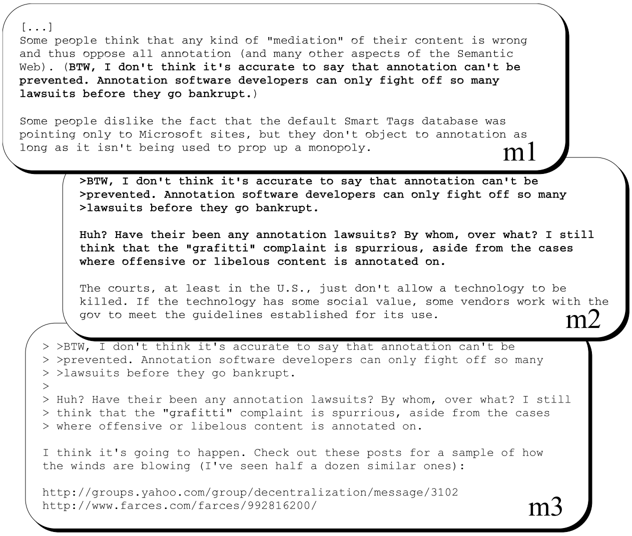
Abstract. Annotation-based discussions are an important concept for today's digital libraries and those of the future, containing additional information to and about the content managed in the digital library. To gain access to this valuable information, discussion search is concerned with retrieving relevant annotations and comments w.r.t. a given query, making it an important means to satisfy users' information needs. Discussion search methods can make use of a variety of context information given by the structure of discussion threads. In this paper, we present and evaluate discussion search approaches which exploit quotations in different roles as highlight and context quotations, applying two different strategies, knowledge and relevance augmentation. Evaluation shows the suitability of these augmentation strategies for the task at hand; especially knowledge augmentation using both highlight and context quotations boosts retrieval effectiveness w.r.t. the given baseline.

1 Introduction

Annotation-based discussions have been identified as an important concept for future digital libraries, supporting collaboration between users [3]. With annotations, a user can comment on the material at hand and others' annotations. As an example for an existing system, the COLLATE prototype uses nested public annotations as a building block for collaborative discussion in a community of scientists, with the purpose of interpreting the digital material at hand [13]. Other examples are web-based newswire systems like ZDNet News[1] which allow users to annotate published articles and other users' comments. In each of these systems, users can change their role from a passive reader to an active content provider. Stored discussion threads can be a helpful source for satisfying users' information needs: On the one hand, annotations can be exploited as auxiliary objects for *document search*, and on the other hand they are retrieval targets themselves in *discussion search*. It becomes clear that discussion search is an important means for uncovering valuable knowledge in information systems such as digital libraries.

[1] http://news.zdnet.com/

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 279–290, 2006.



```
[...]
Some people think that any kind of "mediation" of their content is wrong
and thus oppose all annotation (and many other aspects of the Semantic
Web). (BTW, I don't think it's accurate to say that annotation can't be
prevented. Annotation software developers can only fight off so many
lawsuits before they go bankrupt.)

Some people dislike the fact that the default Smart Tags database was
pointing only to Microsoft sites, but they don't object to annotation as
long as it isn't being used to prop up a monopoly.
```

m1

```
>BTW, I don't think it's accurate to say that annotation can't be
>prevented. Annotation software developers can only fight off so many
>lawsuits before they go bankrupt.

Huh? Have their been any annotation lawsuits? By whom, over what? I still
think that the "grafitti" complaint is spurrious, aside from the cases
where offensive or libelous content is annotated on.

The courts, at least in the U.S., just don't allow a technology to be
killed. If the technology has some social value, some vendors work with the
gov to meet the guidelines established for its use.
```

m2

```
> >BTW, I don't think it's accurate to say that annotation can't be
> >prevented. Annotation software developers can only fight off so many
> >lawsuits before they go bankrupt.
>
> Huh? Have their been any annotation lawsuits? By whom, over what? I still
> think that the "grafitti" complaint is spurrious, aside from the cases
> where offensive or libelous content is annotated on.

I think it's going to happen. Check out these posts for a sample of how
the winds are blowing (I've seen half a dozen similar ones):

http://groups.yahoo.com/group/decentralization/message/3102
http://www.farces.com/farces/992816200/
```

m3

Fig. 1. A discussion thread

In this paper we present our discussion search approaches based on strategies called *knowledge* and *relevance augmentation*, respectively. The methods and results reported here continue the work and preliminary evaluation introduced in [4, 5]. In the next section we briefly present the test collection and our view on emails as annotations. We then introduce possible discussion search approaches in a probabilistic, logic-oriented framework. Subsequently we describe our experiments and discuss their results. We conclude after presenting some related work.

2 The Annotation View on Emails

In order to evaluate our discussion search approaches discussed below, we had to find a suitable test collection. Due to the lack of a "real" digital library testbed containing annotation threads, we participated in last year's TREC Enterprise Track[2] in the discussion search task, where relevant emails had to be found [4]. The collection consists of 174,307 emails from several W3C discussion lists. Figure 1 shows an example excerpt of a discussion thread. Email replies usually consist of two different parts; the *quotations*, which are passages from the original text, and the *new part* containing the actual comments (annotations) of the email author. Quotations are usually prefixed by quotation characters like '>'; combinations of them determine the quotation depth. Quotations are thus the document fragments a comment belongs to. As an example, in m2 the comment "Huh?...established for its use" belongs to the fragment "BTW...go bankrupt"

[2] `http://www.ins.cwi.nl/projects/trec-ent/`

of m1. Applying the distinction between new parts and quotations as well as the thread structure extracted from email headers, we can transform email discussion threads into annotation threads with fragments (determined by quotations) as annotation targets. Due to the fact that whole new parts of emails were the primary target of the discussions search task, we applied one simplification. All quotations and all new parts of an email were merged, so that each email now consisted of one (merged) new part and at most one (merged) quotation part.

3 Discussion Search Approaches

We implement our retrieval functions in predicate logic, in particular probabilistic Datalog (pDatalog). We will briefly introduce pDatalog before discussing our retrieval approaches.

3.1 Probabilistic Datalog

pDatalog [7] is a probabilistic variant of predicate logic. Similar to Prolog, its syntax consists of variables, constants, predicates and Horn clauses. Capital letters denote variables. Probabilities can be assigned to facts. Consider the following example program:

```
0.7 about(d1,"databases").   0.5 about(d1,"retrieval").
retrieve(D) :- about(D,"databases").
retrieve(D) :- about(D,"retrieval").
```

Probabilistic facts model extensional knowledge. The `about` predicate says that `d1` is about 'databases' with 0.7 probability and about 'retrieval' with 0.5 probability. Rules model intensional knowledge, from which new probabilistic facts are derived. The rule `retrieve` means that a document should be retrieved when it is about 'databases' or 'retrieval'. With the given facts and rules, pDatalog would now calculate the retrieval status values of a document d w.r.t. the retrieval function `retrieve` as a combination of probabilistic evidence. In particular, if $e_1, \ldots, e_n$ are joint independent events, pDatalog computes

$$P(e_1 \wedge \ldots \wedge e_n) = P(e_1) \cdot \ldots \cdot P(e_n) \tag{1}$$

$$P(e_1 \vee \ldots \vee e_n) = \bigoplus_{i=1}^{n} P(e_i) = \sum_{i=1}^{n} (-1)^{i-1} \left(\sum_{\substack{1 \le j_1 < \\ \ldots < j_i \le n}} P(e_{j_1} \wedge \ldots \wedge e_{j_i}) \right) \tag{2}$$

For our example document `d1`, pDatalog would calculate

$$\begin{aligned} P(\texttt{retrieve(d1)}) &= P(\texttt{about(d1,"databases")} \vee \texttt{about(d1,"retrieval")}) \\ &= 0.7 + 0.5 - 0.7 \cdot 0.5 = 0.85 \end{aligned}$$

3.2 Simple Content-Based Approach

In this baseline approach, we do not apply any context at all for discussion search. Each document only contains new parts of an email, stripping all quotations. The approach can be expressed with the following datalog rules:

```
wqterm(T) :- qterm(T) & termspace(T).
about(T,D) :- term(T,D).
retrieve(D) :- wqterm(T) & about(T,D).
```

The `qterm` predicate contains the query terms (after stemming and stopword elimination). `termspace` contains the termspace, here regarding only the new part of an email as a document. `termspace` thus contains all terms appearing in new parts of emails. For each term t in the termspace, it is $P(\texttt{termspace(t)}) = P(t)$ which is interpreted as an intuitive measure of the probability of t being informative. $P(t)$ can be estimated based on the inverse document frequency of t, $\mathit{idf}(t) = -\log(\mathit{df}(t)/\mathit{numdoc})$, with $\mathit{df}(t)$ as the number of documents in which t appears and *numdoc* as the number of documents in the collection, and

$$P(t) \approx \frac{\mathit{idf}(t)}{\mathit{maxidf}} \tag{3}$$

with *maxidf* being the maximum inverse document frequency. The `wqterm` rule states that we weight a query term t according to $P(t)$. `term` relates terms to the documents they appear in. For each term t in document d, $P(\texttt{term(t,d)}) = P(t|d)$, the probability that we observe term t given document d. $P(t|d)$ is estimated as

$$P(t|d) \approx \frac{tf(t,d)}{avgtf(d) + tf(t,d)} \tag{4}$$

where $tf(t,d)$ is the frequency of term t in document d and $avgtf(d)$ is the average term frequency of d, calculated as $avgtf(d) = \sum_{t \in d^T} tf(t,d)/|d^T|$ with d^T being the document representation of d (i.e. the bag of words of d). We say that a document is about a term if the term appears in the document; this is modeled with the `about` rule. The `retrieve` rule is our actual retrieval function. A document should be retrieved if it contains at least one query term. The retrieval status value of d is determined by $P(\texttt{retrieve(d)})$ which in turn depends on query and document-term weights, as described above. The result list presented to the user ranks documents according to descending retrieval status values.

3.3 Context Quotations

In the last subsection we were only considering new parts of email messages for retrieval. However, in an email discussion thread, we have the information about the targets that a comment addresses, given by the quotations. Quotations are an important source for determining what a new part is about, as can be seen in message m3 in our example in Figure 1. If we only consider the new part of the message, as we do in the approach described above, the system could not infer that this part is actually about "annotation lawsuits". m3 would not be retrieved for such a query, although it would be relevant. Quotations thus establish an important context for the new parts of messages; quotations are referred to as *context quotations* when regarding them as such a kind of context for new parts. We will now introduce our idea of exploiting context quotations for discussion search, beginning with the obvious choice, merging quotations and new parts by not distinguishing between them in email messages.

Merging Quotations and New Parts. This simple approach sees whole emails as a document (instead of only new parts as in Subsection 3.2). Any further parsing of email messages to distinguish between quotations and new parts is not required here; all terms in a message are regarded as belonging to the corresponding document. In an annotation scenario, this is similar to the case where all annotation targets are merged with their respective annotation to form a new document. We apply exactly the same predicates and rules like those discussed in Subsection 3.2, except that the estimations of $P(t)$ and $P(t|d)$, respectively, are now based on the view of a document being a full email message (resulting in different values for term and document frequencies).

Knowledge Augmentation. While in the last approach context quotations were merged with new parts, the approaches discussed next regard context quotations as separate, virtual documents. Thus, from a message m, two new documents are created: d_m, containing the new part of m, and $quot_m$, containing m's quotations. $quot_m$ is regarded as "virtual" since it is not to be retrieved, but serves as an auxiliary document to determine the relevance of d_m. Furthermore, each virtual document does not contribute to the document frequency of a term. If a term t appears in both $quot_m$ and d_m, only its appearance in d_m is counted and used in Equation 3.

We introduce a new predicate `quotedterm(t,d)` which says that the term t appears in the quotation *quot* belonging to document d. It is $P($`quotedterm(t,d)`$) = P(t|quot)$, and the latter probability is estimated with Equation 4. We apply a *knowledge augmentation* approach by extending our `about` rule to

```
about(T,D) :- term(T,D).
about(T,D) :- acc("quotation") & quotedterm(T,D).
```

where `acc("quotation")` describes the event that a quotation is actually accessed when reading the unquoted part. $P($`acc("quotation")`$)$ is thus the probability that a quotation is considered. By extending the `about` rule like this, we augment our knowledge of what a new part is about with the knowledge of what the quotation is about. In this extended context, new terms are introduced which appear in quotations only, and the probability that a document is about a term is raised according to Equations 1 and 2 if we also observe this term in the quotation. The analogy to the real world is that if a user reads the new part first and then the corresponding quotation, she augments her knowledge of what the new part is about. The `wqterm` and the `retrieve` rules are the same as before.

Relevance Augmentation. Another augmentation strategy we are going to evaluate is *relevance augmentation*. Here, we augment the knowledge that a new part is relevant with the knowledge that its corresponding quotation part is relevant. The idea is that we infer to a certain degree the relevance of a new part with the relevance of its quotation part. This context-based relevance decision is performed by the system in two steps. First, the relevance of documents and context quotations w.r.t. the query is determined:

```
rel(D) :- wqterm(T) & about(T,D).
quot_rel(D) :- wqterm(T) & quotedterm(T,D).
```

In the second step, this knowledge is combined, taking into account the probability that we actually access the quotation:

```
retrieve(D) :- rel(D).
retrieve(D) :- acc("quotation") & quot_rel(D).
```

(`wqterm` and `about` are the same as in Section 3.2).

3.4 Highlight Quotations

When a user annotates a (part of a) document, it is assumed that she found it interesting enough to react to it. This means the annotation target is implicitly highlighted and considered important by the annotation author, reaching a kind of *n-way consensus* [9] of the significance of this part if n persons used it as annotation target. Examining the quotations and the quotation levels of emails, we can identify such highlighted parts of previous messages. A highlight quotation of a message m in another message m' is the part of m which is quoted by m', where m' is a (direct or indirect) successor of m in the discussion thread. Consider the following simple example with 3 messages:

```
m1: line1.1              m2:  > line1.2              m3: >> line1.3
    line1.2                   > line1.3                   > line2.1
    line1.3                   line2.1                     line3.1
```

`m1` consists of 3 lines (`line1.1` - `line1.3`). `m2` quotes two of these lines, `line1.2` and `line1.3`. `m3` quotes a line from `m1` (`line1.3`) and from `m2` (`line2.1`). The quotation in `m2` containing `line1.2` and `line1.3` is a highlight quotation of `m1`. Our claim is that `line1.2` and `line1.3` are important due to the fact that they are quoted; `line1.3` seems to be even more important since it is quoted in `m3` as well. For an email message, we create a highlight quotation virtual document from each quotation containing a fragment of this email message. In our example we would create two highlight quotation virtual documents for `m1`: `high_m1-m2` consists of `line1.2` and `line1.3` (the part of `m1` quoted in `m2`), and `high_m1-m3` contains `line1.3` (the part of `m1` quoted in `m3`). For `m2`, one virtual document is created (`high_m2-m3` containing `line2.1`). We use highlight quotation virtual documents as a context for retrieval by performing knowledge and relevance augmentation again.

Knowledge Augmentation. To add highlight quotations, we introduce a new predicate `highlightterm(t,d,high)` where t is a term, d a document and $high$ the highlight quotation where t appears. It is $P(\texttt{highlightterm(t,d,high)}) = P(t|high)$, again estimated with Equation 4. *Knowledge augmentation* is applied by extending the `about` rule:

```
about(T,D) :- term(T,D).
about(T,D) :- acc("highlight") & highlightterm(T,D,H).
```

$P(\texttt{acc("highlight")})$ is the probability that we actually consider (access) highlight quotations. A short note on the evaluation of `highlightterm(T,D,H)` follows. In the second `about` rule, the variable H is free. For a possible valuation `D=d` and `T=t` to determine `about(t,d)`, pDatalog substitutes `highlightterm(t,d,H)` with a disjunction containing all possible values `H` can take. In our example above, let the term 'developers' appear in `line1.3`. Now, with `T="developers"` and `D=m1`, `highlightterm("developers",m1,H)` would resolve to `highlightterm("developers",m1,high_m1-m2)` $\vee$ `highlightterm("developers",m1,high_m1-m3)`. The probability of this disjunction is calculated and multiplied with $P(\texttt{acc("highlight")})$ to gain a probability for the second `about` rule. `wqterm` and `retrieve` are the same as in Section 3.2 here.

Relevance Augmentation. Relevance augmentation with highlight quotations is quite straightforward. Again, we need two steps;

```
rel(D) :- wqterm(T) & about(T,D).
high_rel(D,H) :- wqterm(T) & highlightterm(T,D,H).
```

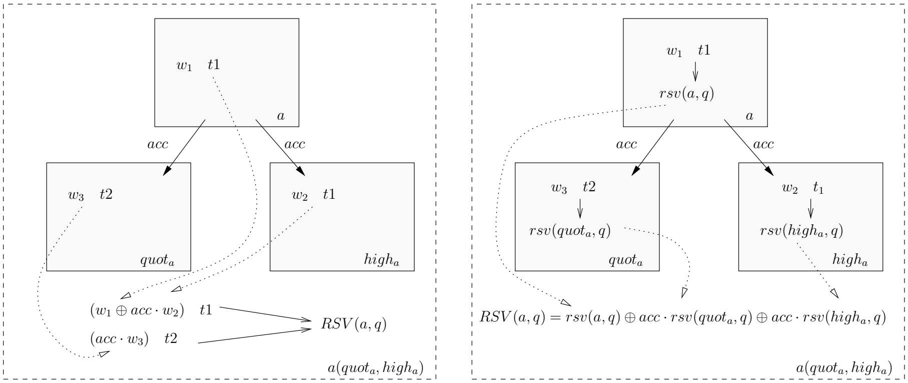
determines the relevance of documents and highlight quotations, and

```
retrieve(D) :- rel(D).
retrieve(D) :- acc("highlight") & high_rel(D,H).
```

combines this evidence in the actual retrieval rules. `wqterm` and `about` are the same as in Section 3.2.

3.5 Combination

We also conducted experiments where we combined the evidence gained from highlight and context quotations. For knowledge augmentation, we combined the corresponding `about` rules introduced in Sections 3.3 and 3.4 with `wqterm` and `retrieve` identical as in Section 3.2. For relevance augmentation, we combined the `rel`, `high_rel`, `quot_rel` and `retrieve` rules in Sections 3.3 and 3.4, respectively, with `wqterm` and `about` as before in Section 3.2.



a) Knowledge augmentation

b) Relevance augmentation

Fig. 2. Knowledge and relevance augmentation

3.6 Non-probabilistic Formulation

The knowledge and relevance augmentation strategies are not bound to a probabilistic, logic-based formulation like the one we presented above with pDatalog. Consider the example in Fig. 2. Here we can see an example annotation a with a corresponding highlight quotation document $high_a$ and a context quotation document $quot_a$. acc models the probability that $high_a$ or $quot_a$, respectively, are accessed from a. With knowledge augmentation, the term weights (w_1 and w_2 for $t1$ and w_3 for $t2$) are propagated to the supercontext $a(quot_a, high_a)$ according to the access probability. The operator $\oplus$ combines the weights from the subcontexts in the supercontext; $\oplus$ can be a simple sum operator, or, as it is the case with pDatalog, formulated with the inclusion-exclusion formula in Equation 2. The calculated new term weights for $t1$ and $t2$ are then used to compute the final retrieval status value $RSV(a,q)$ of a w.r.t. the query q. When applying relevance augmentation, we first calculate a local retrieval status value $rsv(a,q)$, $rsv(quot_a,q)$ and $rsv(high_a,q)$, respectively, for the subcontexts; these values are again combined in the supercontext $a(quot_a, high_a)$ with the $\oplus$ operator in order to compute the final retrieval status value $RSV(a,q)$.

4 Experiments and Results

The main goal of our experiments was to answer the question: can relevance or knowledge augmentation increase retrieval effectiveness, and which strategy should be preferred? Whereas for knowledge augmentation the first question has already been answered [4], we have as yet not conducted any experiments for relevance augmentation. We also provide the results of further runs for knowledge augmentation, applying different values for $P($`acc("quotation")`$)$ and $P($`acc("highlight")`$)$, respectively. For both probabilities, we used global values ranging from 0.1 to 1.0, in steps of 0.1[3]. For our experiments, we used the W3C email lists described in Section 2 with 59 distinct queries. Topics and relevance judgements were given by the participants of the TREC 2005 Enterprise track. All runs were performed using HySpirit[4], a pDatalog implementation. Table 1 briefly describes the experiments and their settings.

Table 1. Description of experiments

Experiment	**Parameters**	**Description**
baseline		The baseline, only new parts.
merged		Merged quotations and new parts
qknow-x	$P($`acc("quotation")`$) = x$	Knowledge augmentation with context quotations
qrel-x	$P($`acc("quotation")`$) = x$	Relevance augmentation with context quotations
hknow-x	$P($`acc("highlight")`$) = x$	Knowledge augmentation with highlight quotations
hrel-x	$P($`acc("highlight")`$) = x$	Relevance augmentation with highlight quotations
cknow-x-y	$P($`acc("highlight")`$) = x$ $P($`acc("quotation")`$) = y$	Knowledge augmentation with highlight and context quotations
crel-x-y	$P($`acc("highlight")`$) = x$ $P($`acc("quotation")`$) = y$	Relevance augmentation with highlight and context quotations

Some selected results of our experiments are presented in Table 2, where we show the mean average precision and the precision at selected numbers of documents retrieved. The latter values are important user-oriented ones: users tend to browse through the first 20 or even 30 top-ranked documents in a result list, but usually do not go deeper in the ranking. The other runs not presented here did not gain better results or considerably new insights. From the results we can see that both relevance and knowledge augmentation improve retrieval effectiveness: there are slight improvements with highlight quotations, and larger improvements with context quotations. To our surprise, the experiment with merged context quotations and new parts gains worse results than the baseline. The combination of highlight and context quotations further improves retrieval effectiveness. So we see that creating separate virtual documents from highlight and context quotations and linking them with a certain access probability to their corresponding document seems to be worth the effort. Regarding knowledge vs. relevance augmentation, the results clearly show that knowledge augmentation is to be preferred over relevance augmentation. In the case of context quotations, knowledge augmentation can possibly handle the vocabulary problem better

[3] We bear in mind that this is only a preliminary solution; more advanced ones might take evidence from the thread structure or given by users' preferences to estimate the access probability.

[4] `http://qmir.dcs.qmul.ac.uk/hyspirit.php`

(when query terms do not appear in the new part, but in the quotation), but the exact reasons are not yet clear and subject to further investigation. Figure 3 shows the interpolated recall-precision averages of selected runs.

Table 2. Mean average precision and precision at 5, 10, 20 and 30 documents retrieved for some selected runs. Best results are printed in bold.

Run	**MAP**	**P@5**	**P@10**	**P@20**	**P@30**
baseline	0.2599	0.4441	0.4103	0.3695	0.3220
merged	0.2565	0.4678	0.3966	0.3458	0.3068
qknow-0.7	0.3162	0.5220	0.4678	**0.3983**	**0.3537**
qknow-0.8	0.3145	0.5186	0.4712	0.3915	0.3503
hknow-0.7	0.2784	0.5085	0.4390	0.3534	0.3124
hknow-0.4	0.2767	0.4542	0.4356	0.3737	0.3266
hknow-0.3	0.2726	0.4915	0.4458	0.3720	0.3260
qrel-0.4	0.2957	0.4746	0.4390	0.3847	0.3401
hrel-0.1	0.2669	0.4780	0.4288	0.3636	0.3192
cknow-0.7-0.8	**0.3298**	**0.5492**	0.4746	0.3890	0.3475
cknow-0.3-0.7	0.3252	0.5458	**0.4881**	0.3975	0.3520
crel-0.1-0.4	0.3024	0.4814	0.4424	0.3831	0.3367

5 Related Work

The studies performed by the Marshall group (see, e.g., [9, 10]) contain many results and conclusions relevant for designers of annotation systems, which have

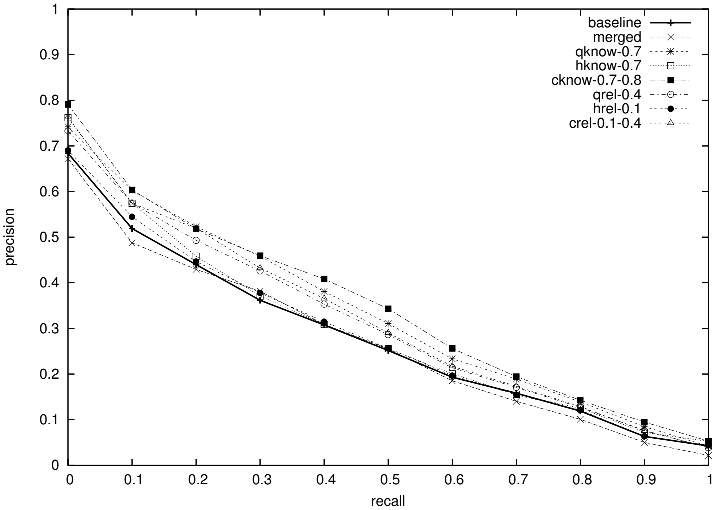


Fig. 3. Interpolated recall-precision graph of selected runs

a strong impact on our work. The studies reported in Shipman *et al.* [12] focus on the identification of high-value annotations in order to find useful passages in a text. Agosti *et al.* examine annotations from a syntactic, semantic and pragmatic view [2].

There are several approaches for annotation-based information retrieval and discussion search. A relevance feedback approach where only highlighted terms instead of whole documents are considered is reported to be successful [8]. [1] reports on an approach where evidence coming from documents and the elements in the annotation hypertext is combined using a data fusioning approach. Xi *et al.* evaluate a feature-based approach for discussion search in [15]; their results show an increase in retrieval effectiveness when using the thread context. The proceedings of the TREC 2005 conference contain many other evaluations of discussion search approaches [14]. The idea of knowledge augmentation has its roots in structured document retrieval and is discussed thoroughly in [11].

6 Conclusion

In this paper we presented some approaches for discussion search and their evaluation, using quotations in a special role as context and highlight quotations, respectively. Based on probabilistic datalog, we applied two strategies, knowledge and relevance augmentation. The results indicate that a knowledge augmentation strategy combining highlight and context quotations is preferable. Knowledge augmentation has another benefit: it is query independent to a certain degree, meaning that $P(\texttt{about(t,d)})$ may be calculated offline as a post-indexing step, whereas the relevance augmentation strategy can only be applied during query processing. Based on the promising results gained so far, we proposed a probabilistic, object-oriented logical framework for annotation-based retrieval called POLAR in [5].

Future work will concentrate on further evaluation and discussion of our augmentation strategies using context and highlight quotations. As a third source of evidence, the content of annotations made to another annotation could also be used for augmentation, as discussed for relevance augmentation in [6].

References

1. Maristella Agosti and Nicola Ferro. Annotations as context for searching documents. In Fabio Crestani and Ian Ruthven, editors, *Information Context: Nature, Impact, and Role: 5th International Conference on Conceptions of Library and Information Sciences, CoLIS 2005*, volume 3507 of *Lecture Notes in Computer Science*, pages 155–170, Heidelberg et al., June 2005. Springer.
2. Maristella Agosti, Nicola Ferro, Ingo Frommholz, and Ulrich Thiel. Annotations in digital libraries and collaboratories – facets, models and usage. In Rachel Heery and Liz Lyon, editors, *Research and Advanced Technology for Digital Libraries. Proc. European Conference on Digital Libraries (ECDL 2004)*, Lecture Notes in Computer Science, pages 244–255, Heidelberg et al., 2004. Springer.

3. Alberto Del Bimbo, Stefan Gradmann, and Yannis Ioannidis. Future research directions – 3rd DELOS brainstorming workshop report, DELOS Network of Excellence, July 2004.
4. Ingo Frommholz. Applying the annotation view on messages for discussion search. In Voorhees and Buckland [14].
5. Ingo Frommholz and Norbert Fuhr. Probabilistic, object-oriented logics for annotation-based retrieval in digital libraries. In *Proceedings of JCDL 2006*, 2006. In print.
6. Ingo Frommholz, Ulrich Thiel, and Thomas Kamps. Annotation-based document retrieval with four-valued probabilistic datalog. In Thomas Roelleke and Arjen P. de Vries, editors, *Proceedings of the first SIGIR Workshop on the Integration of Information Retrieval and Databases (WIRD'04)*, pages 31–38, Sheffield, UK, 2004.
7. Norbert Fuhr. Probabilistic Datalog: Implementing logical information retrieval for advanced applications. *Journal of the American Society for Information Science*, 51(2):95–110, 2000.
8. Gene Golovchinsky, Morgan N. Price, and Bill N. Schilit. From reading to retrieval: freeform ink annotations as queries. In *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval*, pages 19–25, New York, 1999. ACM.
9. Catherine C. Marshall. Toward an ecology of hypertext annotation. In *Proceedings of the ninth ACM conference on Hypertext and hypermedia: links, objects, time and space – structure in hypermedia systems*, pages 40–49, 1998.
10. C.C. Marshall and A.J Brush. Exploring the relationship between personal and public annotations. In *JCDL '04: Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 349–357, New York, NY, USA, 2004. ACM Press.
11. Thomas Rölleke. *POOL: Probabilistic Object-Oriented Logical Representation and Retrieval of Complex Objects.* PhD thesis, University of Dortmund, Germany, 1998.
12. Frank Shipman, Morgan Price, Catherine C. Marshall, and Gene Golovchinsky. Identifying useful passages in documents based on annotation patterns. In Panos Constantopoulos and Ingeborg T. Sølvberg, editors, *Research and Advanced Technology for Digital Libraries. Proc. European Conference on Digital Libraries (ECDL 2003)*, Lecture Notes in Computer Science, pages 101–112, Heidelberg et al., 2003. Springer.
13. Ulrich Thiel, Holger Brocks, Ingo Frommholz, Andrea Dirsch-Weigand, Jürgen Keiper, Adelheit Stein, and Erich Neuhold. COLLATE - a collaboratory supporting research on historic european films. *International Journal on Digital Libraries (IJDL)*, 4(1):8–12, 2004.
14. E. M. Voorhees and Lori P. Buckland, editors. *The Fourteenth Text REtrieval Conference (TREC 2005)*, Gaithersburg, MD, USA, 2005. NIST.
15. Wensi Xi, Jesper Lind, and Eric Brill. Learning effective ranking functions for newsgroup search. In Kalervo Järvelin, James Allen, Peter Bruza, and Mark Sanderson, editors, *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 394–401, New York, 2004. ACM.

Designing a User Interface for Interactive Retrieval of Structured Documents — Lessons Learned from the INEX Interactive Track*

Saadia Malik[1], Claus-Peter Klas[1], Norbert Fuhr[1],
Birger Larsen[2], and Anastasios Tombros[3]

[1] University of Duisburg-Essen, Duisburg, Germany
{malik, klas, fuhr}@is.informatik.uni-duisburg.de
[2] Royal School of Library & Information Science, Copenhagen, Denmark
blar@db.dk
[3] Queen Mary, University of London, United Kingdom
tassos@dcs.qmul.ac.uk

Abstract. The interactive track of the Initiative for the Evaluation of XML retrieval (INEX) aims at collecting empirical data about user interaction behaviour and to build methods and algorithms for supporting interactive retrieval in digital library systems containing structured documents. In this paper we discuss and compare the usability aspects of the web-based user interface used in 2004 with the application based user interface implemented with the Daffodil framework in 2005. The results include a validation of the element retrieval approach, successful implementation of the berrypicking model, and that additional clues for facilitating interactive retrieval (e.g. table of contents, indication of entry points, related terms, etc.) are appreciated by users.

1 Introduction

Many of today's DL systems still treat documents as atomic units, providing little support for searching or navigating along the logical structure of documents. With the steadily increasing use of the eXtensible Markup Language (XML), we have a widely adopted standard format for structured documents. Thus, there is now an opportunity for providing better support for structured documents in digital libraries (DLs). Besides supporting navigation, the logical structure of XML has the potential to assist the DL systems in providing more specific results to users by pointing to document elements rather than to whole documents.

Since 2002, the Initiative for the Evaluation of XML Retrieval (INEX) has organised annual evaluation campaigns for researchers in this field. However, little research has been carried out to study user behaviour and to investigate methods supporting interaction in the context of retrieval systems that take advantage of the additional features offered by XML documents.

* This work was funded by the DELOS Network of Excellence on Digital Libraries (EU 6. FP IST, G038-507618).

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 291–302, 2006.

In order to address these issues, an interactive track (iTrack) was added to INEX in 2004. In this paper, we report on the usability issues addressed in the interactive XML retrieval systems that formed the baseline in these tracks in 2004 and 2005 (hereafter called iTrack 04 and iTrack 05). We show how the findings from the first year led to the development of an improved system in 2005, and we report on the user reactions to both systems.

In iTrack 04, the main goal was to study user behaviour with an XML retrieval system and to validate the element retrieval approach. For this, the user interface design was kept simple in order to give a clear picture of element retrieval systems. During iTrack 04 many usability issues arose, and these led to formulating the main hypotheses for iTrack 05. In addition, more elaborate design principles and the berrypicking paradigm [1] were followed for the iTrack 05 interface design.

This paper is structured as follows: Section 2 gives a brief overview of related work. Section 3 describes the evaluation methodology, the user interface and findings of iTrack 04. The description of the iTrack 05 user interface follows in section 4 including the necessary adaptions derived from iTrack 04, the evaluation and findings. The last section presents a comparison of both evaluations and an outlook.

2 Related Work

Classical information retrieval (IR) research has focused on a system-oriented view and taken a simplified view of user behaviour: the user submits a query and then looks through the ranked items one by one. Thus the goal of the system is to rank relevant items at the top of the list. A broader perspective has been taken in interactive IR research, as represented by the TREC interactive tracks [2]. Quite surprisingly, results of these evaluations showed that differences in system performance identified in laboratory experiments are hard to recreate in interactive retrieval. As described in [3], this result is due to users being able to easily identify relevant entries in a list of documents. Thus, cognitive factors should be considered, as well as richer interaction functions, that can enhance user interaction with the system.

Whereas the standard IR model assumes that the user's information need does not change throughout the search process, empirical studies (e.g. [4]) have shown that interactive retrieval consists of a sequence of related queries targeting different aspects of an ever changing information need. For coping with this problem, Bates et al. has proposed the berrypicking model of information seeking, which assumes that the user's need changes while looking at the retrieved documents, thus leading into new unanticipated directions [1]. During the search, users collect relevant items retrieved by different queries (berrypicking).

So far, there has been little work on interactive XML retrieval. Finesilver and Reid describe the setup of a small collection from Shakespeare's plays in XML, followed by a study of end user interaction with the collection [5]. Two interfaces were used: one highlighting the best entry points and the other highlighting the relevant objects.

Some recent efforts have been made within the INEX interactive track [6, 7]. In addition to the baseline systems which are the topic of this paper, Kamps et al. tested a web-based interface that used a hierarchal result presentation with summarisation and visualisation[8], and van Zwol, Spruit and Baas worked with graphical XML query formulation and different result presentation techniques also in a web-based interface [9]. Besides these systems, various techniques for visualisation of structured documents have been proposed in [10] and [11, 7].

3 iTrack 04

3.1 Evaluation Methodology

Document Corpus. The document corpus used was the 500 MB corpus of 12,107 articles from the IEEE Computer Society's journals covering articles from 1995-2002 [12].

Topics. We used content only (CO) topics that refer to document contents. In order to make the tasks comprehensible by other people besides the topic author, it was required to add why and in what context the information need had arisen. Thus the INEX topics are in effect simulated work task situations as developed by Borlund [13]. Four of the 2004 CO topics were used in the study.

Participating sites. The minimum requirement for sites to participate in the iTrack 04 was to provide runs using 8 searchers on the baseline version of the web-based XML retrieval system provided. 10 sites participated in this experiment, with 88 users altogether.

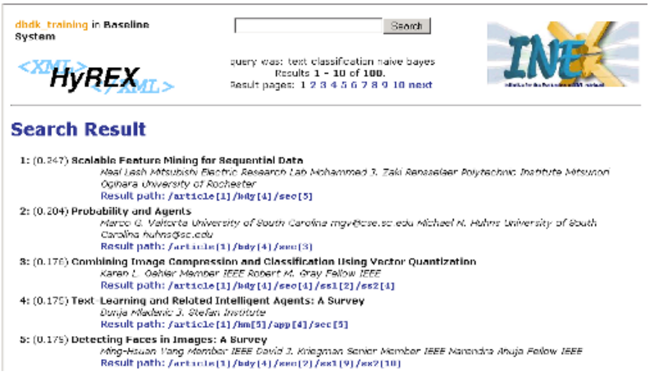
Experimental protocol & data collection. Each searcher worked on one task from each task category. The task was chosen by the searcher and the order of task categories was permuted. The goal for each searcher was to locate sufficient information towards completing a task, in a maximum timeframe of 30 minutes per task.

Searchers had to fill in questionnaires at various points in the study: before the start of the experiment, before each task, after each task, and at the end of the experiment. An informal interview and debriefing of the subjects concluded the experiment. The collected data comprised questionnaires completed by searchers, the logs of searcher interaction with the system, the notes experimenters kept during the sessions and the informal feedback provided by searchers at the end of the sessions.

3.2 User Interface

The user interface in iTrack 04 was a browser-based frontend connecting to the HyREX retrieval engine [14, 15].

In response to a user query, the system presented a ranked list of XML elements including title and author of the document in which the element occurred.



dbdk_training in Baseline System

query was: text classification naive bayes
Results 1 - 10 of 100.
Result pages: 1 2 3 4 5 6 7 8 9 10 next

Search Result

1: (0.247) Scalable Feature Mining for Sequential Data
Neal Lesh Mitsubishi Electric Research Lab Mohammed J. Zaki Rensselaer Polytechnic Institute Mitsunori Ogihara University of Rochester
Result path: /article[1]/bdy[4]/sec[5]

2: (0.204) Probability and Agents
Marco G. Valtorta University of South Carolina mgv@cse.sc.edu Michael N. Huhns University of South Carolina huhns@sc.edu
Result path: /article[1]/bdy[4]/sec[3]

3: (0.176) Combining Image Compression and Classification Using Vector Quantization
Karen L. Oehler Member IEEE Robert M. Gray Fellow IEEE
Result path: /article[1]/bdy[4]/sec[4]/ss1[2]/ss2[4]

4: (0.175) Text-Learning and Related Intelligent Agents: A Survey
Dunja Mladenic J. Stefan Institute
Result path: /article[1]/bm[5]/app[4]/sec[5]

5: (0.175) Detecting Faces in Images: A Survey
Ming-Hsuan Yang Member IEEE David J. Kriegman Senior Member IEEE Narendra Ahuja Fellow IEEE
Result path: /article[1]/bdy[4]/sec[2]/ss1[9]/ss2[10]

Fig. 1. iTrack 04: Query form and resultlist

In addition, a retrieval score expressing the similarity of the element to the query and the path to the element was shown in form of an result path expression (see Figure 1). The searcher could scroll through the resultlist and access element details by clicking on the result path. This would open a new window displaying this element.

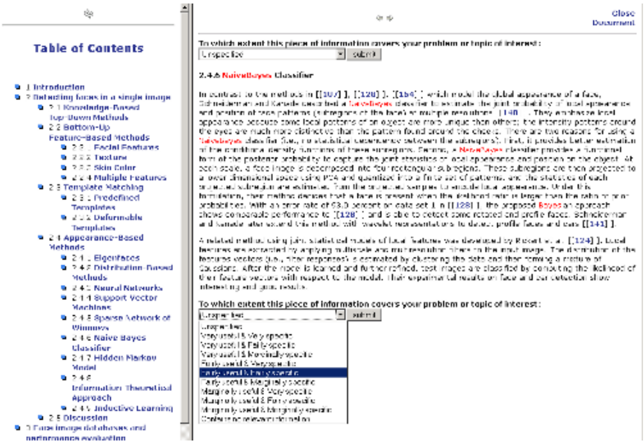


Table of Contents

To which extent this piece of information covers your problem or topic of interest:
Unspecified

2.4.6 NaiveBayes Classifier

In contrast to the methods in [[107]], [[128]], [[154]] which model the global appearance of a face, Schneiderman and Kanade described a NaiveBayes classifier to estimate the joint probability of local appearance and position of face patterns (subregions of the face) at multiple resolutions [[140]]. They emphasize local appearance because some local patterns of an object are more unique than others; the intensity patterns around the eyes are much more distinctive than the pattern found around the cheeks. There are two reasons for using a NaiveBayes classifier (i.e., no statistical dependency between the subregions). First, it provides better estimation of the conditional density functions of these subregions. Second, a NaiveBayes classifier provides a functional form of the posterior probability to capture the joint statistics of local appearance and position on the object. At each scale, a face image is decomposed into four rectangular subregions. These subregions are then projected to a lower dimensional space using PCA and quantized into a finite set of patterns, and the statistics of each projected subregion are estimated from the projected samples to encode local appearance. Under this formulation, their method decides that a face is present when the likelihood ratio is larger than the ratio of prior probabilities. With an error rate of 93.0 percent on data set 1 in [[128]], the proposed Bayesian approach shows comparable performance to [[128]] and is able to detect some rotated and profile faces. Schneiderman and Kanade later extend this method with wavelet representations to detect profile faces and cars [[141]].

A related method using joint statistical models of local features was developed by Rickert et al. [[124]]. Local features are extracted by applying multiscale and multiresolution filters to the input image. The distribution of the features vectors (i.e., filter responses) is estimated by clustering the data and then forming a mixture of Gaussians. After the model is learned and further refined, test images are classified by computing the likelihood of their feature vectors with respect to the model. Their experimental results on face and car detection show interesting and good results.

To which extent this piece of information covers your problem or topic of interest:
Unspecified

- Unspecified
- Very useful & Very specific
- Very useful & Fairly specific
- Very useful & Marginally specific
- Fairly useful & Very specific
- Fairly useful & Fairly specific
- Fairly useful & Marginally specific
- Marginally useful & Very specific
- Marginally useful & Fairly specific
- Marginally useful & Marginally specific
- Contains no relevant information

Fig. 2. iTrack 04: Detail view of an element

The detailed element view is depicted in Figure 2. The content of the selected element was presented on the right hand side. The left hand part of the view showed the table of contents (TOC) of the whole document. Searchers could access other elements within the same document, either by clicking on entries in the TOC, or by using the Next and Previous buttons (top of right hand part). A relevance assessment for each viewed element could be given on two dimensions of relevance: how useful and how specific the element was in relation to the task. These dimensions corresponded to the relevance dimensions of the main ad-hoc track of INEX in an attempt to ensure comparability of the results of the two tracks. Each dimension had three grades of relevance, and ten possible combinations of these dimensions could be given in a drop down list as shown in Figure 2.

3.3 Findings

The main findings based on the log and questionnaires are reported in [16]. Here, only the findings related to the usability of the baseline system are discussed. We analysed the questionnaire and interview data to investigate these issues. Most questionnaire questions were answered on a 5-point scale, which we have analysed statistically.

The overall opinion of the participants about the baseline system was recorded in the final questionnaire which they filled after the completion of both tasks. Users were asked to rate the different features of the system on the scale of 1 to 5, where 1 stood for 'Not at all', 3 'Somewhat' and 5 for 'Extremely'. The results are summarised in Table 3.

In addition to these ratings, users were asked to comment on the different aspects of the system after the completion of each task and after the completion of the experiment. Example questions were:

- *In what ways (if any) did you find the system interface useful in the task?*
- *In what ways (if any) did you find the system interface not useful in the task?*
- *What did you like about the search system? What did you dislike about the system?* and
- *Do you have any general comments?*

The analysis of the most frequent comments are presented in the following sections. Table 1 summarises the positive and Table 2 the negative results.

Element overlap. One of the critical issues of element retrieval is the possible retrieval of overlapping result elements, i.e. components from the same document where one includes the other (due to the hierarchic structure of XML documents). Typically these elements are shown at non-adjacent ranks in the hit list. In our case, the HyREX retrieval engine did not take care of overlapping elements and thus searchers frequently ended up accessing elements of the same document at different points in time and at different result ranks.

Data from both the system logs and the questionnaires showed that searchers found the presence of overlapping elements distracting. By recognising that they had accessed the same document already through a different retrieved element, searchers typically would return to the resultlist and access to another element instead of browsing again within a document visited before. 31 users commented negatively on the element overlap.

Document structure provides context. The presence of the logical structure of the documents alongside the contents of the accessed elements was a

Table 1. Positive responses on system usefulness (iTrack 04, 88 searchers)

System Features	Response Count
Table of contents	66
Keyword highlighting	36
Simple/easy	34
Good results	13
Fast	8
Simple querying	6

Table 2. Negative responses on system usefulness (iTrack 04, 88 searchers)

System Features	Response Count
Overlapping elements	31
Insufficient summary	30
Distinction b/w visited & unvisited	24
Limited query language	22
Poor results	10
Limited collection	9
Slow	9

feature that searchers commented positively on. The table of contents of each document (see Figure 2) seemed to provide sufficient context to searchers in order to decide on the usefulness of the document. 66 users found the TOC of the whole article very useful because it provided easy browsing, navigation, less scrolling or gave a quick overview of which elements might be relevant and which might not be.

Element summaries. The resultlist presentation in the iTrack 04 system did not include any element summarisation. Only the title and authors of the document were displayed in addition to the result path expression of the element and its similarity to the query. As a consequence searchers had little clues available to decide on the usefulness of retrieved elements at this point. 30 users commented on these insufficient clues.

Keyword highlighting. Within the detail presentation of an element, all query terms were highlighted. This feature was very much appreciated, and several users suggested to provide this feature not only at the resultlist level, but also at the table of contents level. 36 users gave positive comments on this feature.

Distinction between visited and unvisited elements. There was no distinction between visited and unvisited elements at the resultlist and detail levels. Thus, a number of times users visited the same elements/documents more than once. 24 users commented negatively on this.

Limited query language. The system did not support sophisticated queries and there was no possibility to use phrases, boolean queries, or to set the preference for terms. 22 users found this an obstacle.

General issues. There are also some more general issues that were commented on. These stated that the multiple windows of the web-interface were somewhat confusing and that the "Result path" shown in the resultlist was mostly meaningless, and with the square brackets, it had a very technical appearance.

iTrack 04 was the first attempt to set up an interactive track for XML retrieval, and there was very little knowledge on which we could build upon when designing the iTrack 04 interface. In contrast, the design of the iTrack 05 interface was based on the expereinces from the previous year. In designing the interface, we aimed at overcoming the main weaknesses of the 2004 interface.

4 iTrack 05

4.1 Evaluation Methodology

The evaluation methodology used in iTrack 05 was similar to the one used in iTrack 04. An extended version of the INEX IEEE document collection was used (now comprising 16819 documents).

This time six topics were selected from the INEX 2005 ad-hoc topics, and modified into simulated work tasks. In addition, searchers were asked to supply two examples of their own information needs. Depending on the coverage in the collection, one of these tasks was selected by the experimenter for the experiment. In total, each searcher performed three tasks. With a total of 11 participating organisations, 76 searchers performed 228 tasks in iTrack 05.

4.2 Desktop-Based System

For iTrack 05 the Daffodil framework was used and extended to meet the functionality of XML retrieval. Daffodil is a front-end to federated, heterogeneous digital libraries. It is aimed at providing strategic support (see [17]) during the information search process and already supports interactive retrieval through integrated high-level search and browse services.

The Daffodil framework consists of two parts, the graphical user interface client and the agent-based backend services (see [18, 19]). The user interface client, implemented in Java, is based on a tool metaphor, where each service is presented by a tool and the tools are integrated among each other.

The interface for iTrack 05 was designed by taking into account the findings of the iTrack 04, the berrypicking model described in section 2 and iconic visulisation techniques for better recall and immediate recognition.

Additions to the Architecture. The base system had to be extended for INEX in order to deal with the highly structured XML data. These extensions affected both the user interface and the corresponding backend services, e.g. connecting the XML search engine.

Query formulation. The problem of limited query language expressiveness was resolved by allowing Boolean queries, in combination with proactive query formulation support [20]. The latter feature recognises syntactic errors and spelling mistakes, and marks these. Besides full-text search, the system now also allowed for searching on metadata fields such as authors, title, year.

For further support during query formulation we added a Daffodil service for suggesting related query terms (based on statistical analysis of a different corpus). While the user specifies her query, a list of possible alternative terms are presented to her. This service follows the berrypicking model because the newly discovered related terms can change the search direction of the user. For easy query reformulation, the drag&drop feature of DAFFODIL could be used to add new query terms from documents or the related term list.

Resultlist presentation. In order to resolve the issues of *overlapping elements* and *element summarisation* identified in iTrack 04, results in the resultlist were now grouped document-wise and hits within documents were presented as possible entry points within the hierarchical document structure. The document metadata information is shown as the top level element, as depicted in Figure 3.

In addition, whenever some element within a document is retrieved, the title of that element is presented as a document entry point, depicted as a clickable folder icon. This change reflected user preference for the TOC view, where titles of elements are displayed.

We also took into account the comments about the retrieval score and the result path expression from iTrack 04. The retrieval score of each retrieved element was now shown in pictorial (as opposed to numerical) form, and result path expressions of elements were removed from the resultlist. The whole resultlist entry was made clickable.

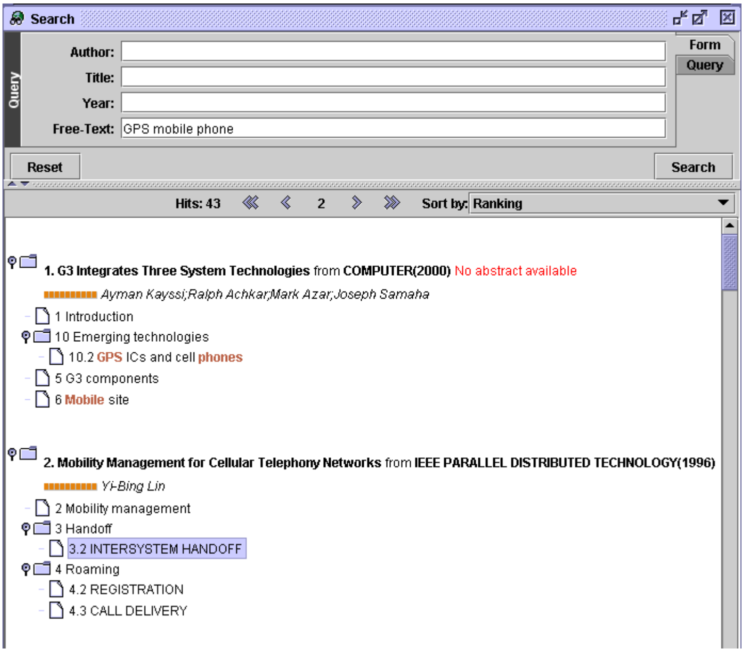
The comments on the distinction between visited and unvisited elements were considered by using an iconic visualisation technique. An eye icon is shown with any resultlist entry that has been visited before. The analogy with the berrypicking model is given here as marking the paths where a user walked to pick only unknown berries, to avoid looking twice at the same information. We also adopted query keyword highlighting at the resultlist level, since searchers appreciated this feature at the detail view level.

Detail view. The main layout of the detail level was kept the same as in iTrack 04, as seen in Figure 4. Some additions were made for supporting document browsing. First, the entry points from the resultlist level are now also highlighted in the detail view. Second, elements already visited are indicated with an iconised eye in the table of contents.

Many participants in iTrack 04 felt that the two-dimensional relevance scale used in these experiments was too complex [21]. For this reason, we moved to a simple 3-point scale, measuring only the usefulness of an element in relation to the searcher's perception of the task: 2 (Relevant), 1 (Partially Relevant), and 0 (Not Relevant). This three grade relevance scale was visualised as shown in Figure 4 (top left hand). The same icons were added to the viewed element when a relevance value was assigned by the user. Here again one more aspect of the berrypicking model analogy was implemented successfully: the user puts the 'good' beeries into her basket, and also can see which berries she has picked before.

4.3 Findings

The analysis was made along the same lines as for iTrack 04. The overall opinion of the participants about the system was recorded in the final questionnaire



Author:

Title:

Year:

Free-Text: GPS mobile phone

Hits: 43 « ‹ 2 › » Sort by: Ranking

1. G3 Integrates Three System Technologies from COMPUTER(2000) No abstract available
Ayman Kayssi;Ralph Achkar;Mark Azar;Joseph Samaha
- 1 Introduction
- 10 Emerging technologies
  - 10.2 GPS ICs and cell phones
- 5 G3 components
- 6 Mobile site

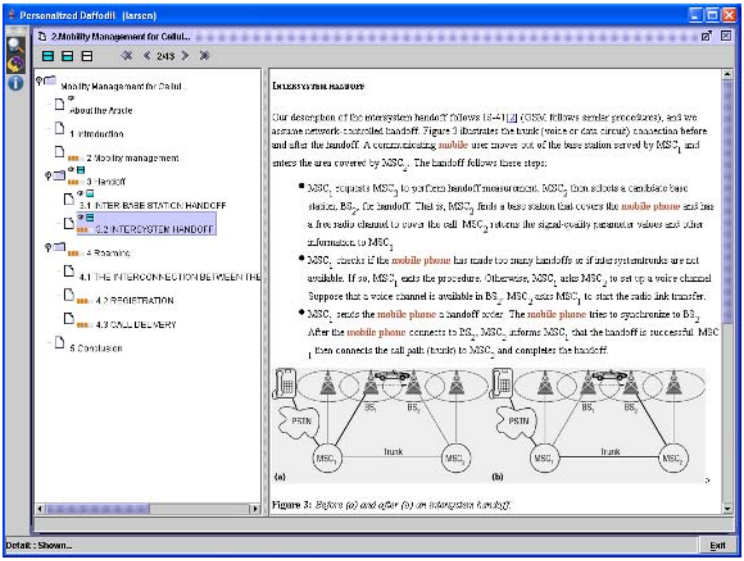
2. Mobility Management for Cellular Telephony Networks from IEEE PARALLEL DISTRIBUTED TECHNOLOGY(1996)
Yi-Bing Lin
- 2 Mobility management
- 3 Handoff
  - 3.2 INTERSYSTEM HANDOFF
- 4 Roaming
  - 4.2 REGISTRATION
  - 4.3 CALL DELIVERY

Fig. 3. iTrack 05: Query form and resultlist

Table 3. Overall opinion about the system on the scale of 1 (Not at all) to 5 (Extremely) in iTrack 04 (88 searchers) & iTrack 05 (76 searchers)

System Features	iTrack 04		iTrack 05	
	μ	σ^2	μ	σ^2
How easy was it to learn to use the system?	4.17	0.6	3.40	0.9
How easy was it to use the system?	3.95	0.7	3.96	0.9
How well did you understand how to use the system?	3.94	0.5	3.84	0.9
How well did the system support you in this task?	-	-	3.13	0.9
How relevant to the task was the information presented to you?	-	-	2.97	1.13
Did you in general find the presentation in the resultlist useful?	-	-	3.35	0.8
Did you find the table of contents in the detail view useful?	-	-	3.72	1.0

that they filled after the completion of all tasks. New questions enquiring about the distinct aspects of the system used in 2005 were added. The results are summarised in Table 3. As can be seen users were in general positive on both systems, and the major difference between the two years was the better learnability of the 2005 system. In addition, there were many informal comments in response to the questions mentioned in section 3.3. We analyse the data in the following paragraphs.



INTERSYSTEM HANDOFF

Our description of the intersystem handoff follows IS-41[2] (GSM follows similar procedures), and we assume network-controlled handoff. Figure 3 illustrates the trunk (voice or data circuit) connection before and after the handoff. A communicating mobile user moves out of the base station served by MSC$_1$ and enters the area covered by MSC$_2$. The handoff follows these steps:

- MSC$_1$ requests MSC$_2$ to perform handoff measurement. MSC$_2$ then selects a candidate base station, BS$_2$, for handoff. That is, MSC$_2$ finds a base station that covers the mobile phone and has a free radio channel to cover the call. MSC$_2$ returns the signal-quality parameter values and other information to MSC$_1$
- MSC$_1$ checks if the mobile phone has made too many handoffs or if intersystemtrunks are not available. If so, MSC$_1$ exits the procedure. Otherwise, MSC$_1$ asks MSC$_2$ to set up a voice channel. Suppose that a voice channel is available in BS$_2$. MSC$_2$ asks MSC$_1$ to start the radio link transfer.
- MSC$_1$ sends the mobile phone a handoff order. The mobile phone tries to synchronize to BS$_2$. After the mobile phone connects to BS$_2$, MSC$_2$ informs MSC$_1$ that the handoff is successful. MSC$_1$ then connects the call path (trunk) to MSC$_2$ and completes the handoff.

Figure 3: *Before (a) and after (b) an intersystem handoff.*

Fig. 4. iTrack 05: Detail view

Resultlist presentation. Presentation of results in a hierarchy is generally found useful. 43 users commented positively on it, whereas 3 users found the information presented insufficient for deciding about relevance or irrelevance. 2 users commented on the inconsistency of the result presentation. This situation occurred when a whole article was retrieved as a hit, with no further elements within this article, 3 users disliked scrolling at the result list level.

Table of contents and query term highlighting. As in iTrack 04, the TOC is found to be extremely useful and 32 users commented positively on it. Query term highlighting in the resultlist and the detail view were also appreciated (22 positive comments).

Related terms. The new functionality of suggesting related query terms was found highly helpful: 29 users found this function useful in their performance of search tasks. There were some cases when the suggested terms either retrieved no documents, or there was no obvious semantic relationship to the query terms. These situations led to negative remarks by 11 searchers.

Awareness in the detail view. The document entry points shown in the resultlist were also displayed in the detail view, 14 users commented positively on it. In addition, icons indicating visited elements and their relevance assessments are shown in the TOC: 3 users found this useful. In addition, 15 users also wanted to have the relevance assessment information in the resultlist.

Retrieval quality. Although the underlying retrieval engine had shown good retrieval results in previous INEX rounds, it produced poor answers for some queries, so 25 users commented negatively on this. A possible reason could be the limited material on the choosen topic of search.

Other Issues. 4 users remarked positively on the interface usefulness and 3 liked the query form. The response time of the system was encountered as being too high, so 35 users comments negatively on it.

Overall, user responses show that the main weaknesses of the iTrack 04 interface have been resolved. In addition, the new features supporting the berrypicking paradigm were appreciated by the users.

5 Conclusion and Outlook

In this article we presented the lessons learned from INEX iTrack 04 to iTrack 05. The analysis of iTrack 04 showed several negative responses to the used web-based interface. The main issues were the overlapping elements presented in a linear resultlist, insufficient summaries to indicate the relevance of an item, the lack of distinction between visited and unvisited items and a limited query language. Also some positive comments were made, e.g., the document structure (TOC) provided sufficient context and was a quick way of locating the interesting information. Keyword highlighting was also found to be helpful in 'catching' information parts that may be relevant to the existing query terms.

These findings were used to shift to an application-based interface. The analysis of iTrack 05 showed that the overlapping elements presentation in a hierarchy can provide sufficient summerisation and context for the decision of relevance or irrelevance. The second major improvement was the addition of design elements based on the berrypicking model [1], which received substantial appreciation. These desgin elements included keyword highlighing, iconic visualisation and provision of related terms.

The most problematic issue with the iTrack 05 system was the responsiveness of the system. This was due to the underlying search engine and inefficiencies within the DAFFODIL message flow. These issues will be worked on for iTrack 06.

Overall, the evaluations showed that interface design adaptation based on the 2004 findings were taken as an improvement. The shift to an application based framework proved to be the right step, as we gained more flexibilty in features besides a web-based framework. In iTrack 06 a major focus will be the efficiency, by replacing the underlying search engine and a tighter integration with the DAFFODIL framework to lower response times.

References

1. Bates, M.J.: The design of browsing and berrypicking techniques for the online search interface. Online Review **13** (1989) 407–424
2. Voorhees, E., Harman, D.: Overview of the eighth Text REtrieval Conference (TREC-8). In: The Eighth Text REtrieval Conference (TREC-8). NIST, Gaithersburg, MD, USA (2000) 1–24

3. Turpin, A.H., Hersh, W.: Why batch and user evaluations do not give the same results. In: Proc. of SIGIR, ACM Press (2001) 225–231
4. O'Day, V.L., Jeffries, R.: Orienting in an information landscape: How information seekers get from here to there. In: Proc. of the INTERCHI '93, IOS Press (1993) 438–445
5. Finesilver, K., Reid, J.: User behaviour in the context of structured documents. In: Proc. of ECIR. (2003) 104–119
6. Larsen, B., Malik, S., Tombros, A.: The interactive track at inex 2005. In: Advances in XML Information Retrieval and Evaluation: Springer, p. 398-410. (Lecture Notes in Computer Science vol. 3977). (2006)
7. Tombros, A., Larsen, B., Malik, S.: The interactive track at inex 2004. In: Advances in XML Information Retrieval: Springer, p. 410-423. (Lecture Notes in Computer Science vol. 3493). (2004)
8. Kamps, J., de Rijke, M., Sigurbjörnsson, B.: University of amsterdam at inex 2005. (In: Advances in XML Information Retrieval and Evaluation: Springer, p. 398-410. (Lecture Notes in Computer Science vol. 3977))
9. van Zwol, R., Spruit, S., Baas, J.: B^3-sdr@inetractive track: User interface design issues. (In: INEX 2005 Workshop Pre-Proceedings)
10. Crestani, F., Vegas, J., de la Fuente, P.: A graphical user interface for the retrieval of hierchically structured documents. Information Processing and Management **40** (2004) 269–289
11. Großjohann, K., Fuhr, N., Effing, D., Kriewel, S.: Query formulation and result visualization for XML retrieval. In: Proceedings ACM SIGIR 2002 Workshop on XML and Information Retrieval. (2002)
12. Gövert, N., Kazai, G.: Overview of the INitiative for the Evaluation of XML retrieval. In: Proc. of INEX workshop. (2003) 1–17
13. Borlund, P.: Evaluation of interactive information retrieval systems. (2000) 276 PhD dissertation.
14. Fuhr, N., Gövert, N., Großjohann, K.: HyREX: Hyper-media retrieval engine for XML. In: Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval. (2002) 449 Demonstration.
15. Gövert, N., Fuhr, N., Abolhassani, M., Großjohann, K.: Content-oriented XML retrieval with HyREX. In: Proc. of INEX workshop. (2003) 26–32
16. Tombros, A., Malik, S., Larsen, B.: Report on the INEX 2004 interactive track. SIGIR Forum **39** (2005)
17. Klas, C.P., Fuhr, N., Schaefer, A.: Evaluating strategic support for information access in the DAFFODIL system. In: Proc. of 8th ECDL. (2004)
18. Fuhr, N., Klas, C.P., Schaefer, A., Mutschke, P.: Daffodil: An integrated desktop for supporting high-level search activities in federated digital libraries. In: Proc. of 6th ECDL, Springer (2002) 597–612
19. Fuhr, N., Gövert, N., Klas, C.P.: An agent-based architecture for supporting high-level search activities in federated digital libraries. In: Proc. of ICADL, Taejon, Korea, KAIST (2000) 247–254
20. Schaefer, A., Jordan, M., Klas, C.P., Fuhr, N.: Active support for query formulation in virtual digital libraries: A case study with DAFFODIL. In: Proc. of 7th ECDL. (2005)
21. Pehcevski, J., Thom, J.A., Vercoustre, A.: Users and assessors in the context of inex: Are relevance dimensions relevant? In: Proc. of INEX Workshop on Element Retrieval Methodology. (2005)

"I Keep Collecting": College Students Build and Utilize Collections in Spite of Breakdowns

Eunyee Koh and Andruid Kerne

Interface Ecology Lab, Center for Study of Digital Libraries, Computer Science Department,
Texas A&M University, College Station, TX 77843, USA
{eunyee, andruid}@cs.tamu.edu

Abstract. As people become more and more involved with digital information, they grow proportionally involved in situated practices of collecting. They put together large sets of information elements. However, their attention to those information elements is limited. They use whatever means are at hand in order to form representations of their collections. They need to keep track of the elements in these collections, so they can use them later. We conducted a study with 20 college students. A major concern for the students during collection building was collection management and utilization, particularly as the size and number of their collections grows. They experienced breakdowns in these processes, yet continued to engage in collecting. They developed strategies such as informal metadata schemas and hierarchical organization to try to cope with their problems. We consider the practices observed, and their implications for the development of tools to support digital collection building and utilization. Collection representations that support cognition, collaboration, and semantic schemas are prescribed.

1 Introduction

Dick is a graduate student in industrial engineering. As he is a research assistant, his work involves writing research papers. He regularly searches for and collects relevant prior work from the internet and digital libraries. He collects articles and URLs on his own computer. He utilizes this collection regularly. Jane is a visualization lab student. She collects many images and pictures for class work such as animation, and also for fun. Some of these are photographs she has taken; some come from the internet. She is also a student worker in the university newspaper. She collects images to support this activity, as well. These examples illustrate the contexts in which students are making collections, and provide a sense of the scope of collections and collecting activities addressed by this paper. We define *collecting* as people's practices of putting together archives of information elements, such as hyperlinks, documents, images, audio, and video, with the intention of creating and supporting meaningful, engaging, and useful experiences.

Due to popularity of digital media devices and the abundance of information on the web, a broad cross-section of society becomes more and more exposed to large numbers of digital documents and media elements. People are confronted with the

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 303 – 314, 2006.

problem of how to keep track of significant elements within the stream of this experience. They begin collecting, and again due to the preponderance of meaningful digital information and media, the collections become larger and larger. This trend is further promulgated by the increasing availability and capacity of inexpensive digital storage devices.

However, a wealth of information creates a poverty of attention [19]. The disparity between the growing amount of information and media that people are collecting in practice, and their fixed amount of attention, is leading to breakdowns in their collecting experiences. According to Winograd and Flores, breakdowns occur when there is a discrepancy between our expectations and actions, and the world [22]. Breakdowns can serve as an opportunity for learning, because they identify important parts of tasks and activities, and can provoke the articulation of new user needs and design requirements. The present research investigates breakdowns in collecting practices.

The study has been conducted with college students. College students tend to be fast movers in the face of ongoing technological transformation. 81% of them go online. Many of them can scarcely imagine what the world was like way back when people weren't always connected to the net, "Always on" [14]. The Pew Internet and American Life Project, reporting on 2054 students from 27 college and university, says that nearly 73% of college students use the internet more than the library, while only 9% said they use the brick and mortar libraries more than the internet for information searching [15]. College students typify the category of *power creators*, which Pew has identified as an important constituency of internet users [16]. Power creators are twice as likely to engage in content creating activities as other internet users [16].

The intention of this research is to develop understanding of current practices and resulting breakdowns in building and utilizing digital collections. We have investigated the practices of college students by interviewing them, and observing the collections that they build. We also gathered quantitative data about collection building and utilization practices. From this understanding, we will infer implications for the design of new tools to support these processes. This paper begins with a review of related work. Next, we describe the study and its participants. The subsequent section presents data and analysis. We conclude by discussing current collection practices and tools, and infer design implications for future research and development.

2 Related Work

Prior studies have investigated the usage of tools for building and utilizing collections in specific media, such as email [3], bookmarks [1][12], and files [4][5]. Some studies have offered classifications of user behavior with various collection tools. Malone identified two fundamental strategies in office management: filing and piling [13], focusing on the organization activities. Whittaker and Sidner [21] observed three email management strategies: frequent filer, spring cleaner, and no filer. Balter [3] extended this classification by dividing the no-filer class into folder-less cleaner and folder-less spring-cleaner, depending on whether items are deleted from the inbox on a daily basis. Abrams *et al.* [1] described four bookmark management strategies: no-filer, creation-time filer, end-of-session filer, and sporadic filer. Barreau and Nardi [5] looked at the types of information manage by users, identifying three types based on lifetime and use: ephemeral, working, and archived. They noted the relative importance of ephemeral/working items retrieved by location-based browsing over archived items and the use of search. However, as the information age matures, it seems that the importance of archiving grows.

While each of the previously mentioned works addresses utilization of a single collection medium, Jones *et al.* conducted a study that traverses collecting practices involving e-mail, images, document addresses (URLs), and documents [12]. They investigated various methods people use in their workplace to organize information for re-use. They found that people differ in their collection building practices according to their job position and their relationship to the information. Their study is similar to the present research in its addressing of multiple collection media, as well as in the number of experimental subjects, and the social proximity of the subjects to the researchers. Boardman *et al.* [7] also collected cross-tool data relating to file, email and web bookmark usage. They found that individuals employ a rich variety of strategies both within and across collection tools, and discuss synergies and differences between tools, to guide the design of tool integration. The data underlined the challenge of the collection tool design by addressing that future design work must take account of the variation in strategies by providing the flexibility to manage different types of information in distinct way. They observed that people usually browse rather than search to find relevant elements in their collections. In addition, they found that the slow-changing nature of hierarchical representations may benefit users by promoting familiarity with the personal information environment. Such familiarity, in turn, supports location-based finding for which users expressed a clearer preference.

The present research focuses on human experiences of collecting and the role of collections across a broad range of meaning-making activities and digital media. Some prior work has addressed particular media, such as web pages or email. Some has focused on well-defined scenarios regarding information filing, finding, and management. This study investigates processes of collection building and utilization across media and tools through open questions about participants' situated practices, in order to discover how they engage in collecting throughout their everyday activities. We use a hybrid data collection approach, in which qualitative data from open questions is augmented by quantitative data about collection building and utilization.

3 Study Description

To investigate power users' collection building and utilizing practices, we performed a study consisting of interviews of 20 college students. The study brought together narrative accounts, interview questionnaires, and examples of their digital collections in order to investigate how they currently build and utilize collections as part of everyday life. Students were informed that they were participating in a study, and that their responses would be recorded, and anonymously recounted in a research paper.

Participants were distributed by gender and academic concentration. Ten students were male and the other ten were female. There were eight undergraduate students and twelve graduate students. Students' majors were diverse, including computer science, visualization, aerospace engineering, statistics, landscape design, industrial engineering, and history. The interviews were conducted with participants at their offices or homes, so they could show artifacts from their personal computers.

The interviews were semi-structured and open-ended. We did not limit the dialogue to our pre-formulated questions. We also did not place any limits on the media type or representational forms of the collections we investigated. Rather, we considered any type of personal collection. We spent 60-90 minutes with each participant to explore the kinds of collections they made, their processes of using and organizing the collections, the collection tools they used, and their overall experiences of collecting.

While conducting the study, the interviewer was guided by an agenda of relevant research questions:

- To what extent do you think intentionally about your needs for collecting digital information prior to actually doing so?
- What activities are involved in your collection building processes?
- How do you feel about spending time through collection making processes?
- How many elements are in your collections?
- Which tool(s) or mechanism(s) do you use to build collections?
- How often do you make / refer to / organize collections?
- What types of inconveniences and breakdowns do you encounter during building and utilizing digital collections?
- What are your strategies for coping with breakdowns in your experiences of building and utilizing collections?
- What are your suggestions for future collection tools?

We recorded and screen-copied examples of collections participants built, and took notes of interviews. After each interview, participants filled out a survey questionnaire.

4 Results

We analyzed the study data in terms of the distribution of activities, significance, type, and quantity of information elements involved, as well as the kinds of mechanisms people used for building and utilizing collections. We also investigated their frequency of involvement in collecting. Quantitative and qualitative data and its analysis will show participants' collection building and utilizing practices and behavior.

4.1 Collection Building and Utilizing

We looked at collection building and utilizing practices in terms of the stance participants brought into the process of collecting, the patterns and expectations that occurred in these processes and the ways in which users perceived success and failure.

Intention and Need

Participants were asked whether they thought about the need for collecting prior to engaging in processes of seeking digital information. All participants expressed awareness of a personal ongoing deliberate intention and need to be involved in collection building and utilizing practices.

Activities and Significance

The participants reported collecting digital media materials that support a range of personal and work-related activities. The personal media included photographs taken by themselves and friends, as well as popular media elements such as music, movie star pictures, and art images. As the subjects were students, their work is learning and research, so the materials here included class notes and research papers. Students whose majors are related to design collect many image files as part of their school work. From this data, we see that the participants' collecting activities are conducted in relationship to the span of significant activities in their lives.

Frequency and Time Period

One hundred percent of participants report that they build and utilize collections regularly. Of these, more than half utilize collections more than one hour per week. In more detail, 18% of participants said that they spend more than one hour per day on collection building; 10% spend one hour per day engaged in the collection process; 27% said that they spend more than one hour per week and less than one hour a day; while 27% spend one hour a week; and 18% of participants spend one hour per month. However, participants do not have a specific time frame scheduled for collection building and utilizing. It is something they do spontaneously, as part of a range of tasks and activities (P3: *"I build and utilize collections regularly, and I engage in this process during spare time and while I am taking rest."*).

Worthwhile or Useless

Participants were asked how they feel about spending time on collection building and utilization. 46% of participants said that they experience the process as meaningful and worthwhile. 18% of participants answered that they find it somewhat meaningful. 9% of participants answered that their experience is neither worthwhile nor useless. 27% of participants said that they experience collecting as rather useless. Those participants who answered rather useless said that they nonetheless continue to engage in the collection building process; they experience it as necessary and meaningful initially, but after a while, their engagement seems to be performed in vain. They said that a collection is not worthwhile if they do not utilize it well, and they seldom utilize most parts of their collections because of the huge volume of collected information.

Collection Types

All participants said that they build image, music, and/or movie collections. The sources of the images are from digital cameras, camera phones, and the internet. Twenty-two percent of participants have 50-100 images in their collection; another 22% keep 100-500 images; while 56% keep more than 5000 image collections. Participants said they mostly obtain music from music downloading services or their friends' collections. Thirty-three percent of participants keep 50-100 music files, 33% keep 100-500 files, and 34% keep more than 500 music files in their collections.

Movie files are obtained through similar means, such as downloading services or creation with a video camera. Twenty-two percent of participants keep 10-50 movie files; another 22% keep 50-100 movie files; another 22% keep 100-500 movie files; while 34% keep more than 500 movie files in their collections.

Participants also collect documents such as Word files and PDF files. 56% of participants keep 100-500 documents; 44% keep more than 500 documents in their collections. They also collect web documents in the form of hyperlinks (URLs). 11% of participants keep 1-10 URLs, 33% keep 10-50 URLs, 45% keep 50-100, and 11% keep 100-500 URLs in their collections. Compared to the other media collections, participants keep fewer URLs, because web documents are easier to search for.

Collection Mechanism

In terms of what is stored, there are three ways to build digital collections: (1) save the files themselves; (2) extract some parts from files and save only those parts; (3) save the location of files. Participants use whatever tools and structures are at hand to build their collections; for example, files, folders, bookmarks, and e-mail.

All participants said that they make file folders for file collections. There are also within-file collections, in which small elements of information from diverse sources are gathered into a single file. Participants said that they used Excel, Word, Photoshop, and Notepad to build this type of within-file collection. They used drawn lines, tables and newline characters (vertical whitespace) to spatially distinguish elements in a within-file collection. When participants save URLs of web pages, they usually use bookmarks, but they also use e-mail, so that those URLs can be utilized from the other computers (P9: *"I am not using bookmarks at all. Instead I keep URLs in my email because I use three computers; my office computer, my home computer, and my laptop, so I can look at important URLs from any of my computers."*).

Levels of Engagement with Collections

We observe that in general, people collect information and media with the intention of later referring to the collected elements for use. Sometimes, they actually get to this process of referring. Further, sometimes, with collections that are important, they take steps to organize the form of the collection. Referring and organizing are aspects of collection utilization.

While participants accessed the internet daily, their activities of selecting elements to add to their collections, referring to the collections, and organizing them occurred less frequently (See Figure 1 Left). The frequency of these activities can be categorized in three tiers. All of the subjects accessed the internet daily. At the same time, 43% of them engaged in collection building and referring on a daily basis, while 36% did so on a weekly basis, and the remaining 21% engaged in such activities monthly. The difference between internet access frequency and collection building/referring frequency was statistically significant [$F(2,26)=3.67$, $p<.01$]. While distribution of the participants' collection building frequency and collection referring frequency were the same, these distributions are independent and do not necessarily refer to the same participant. The third tier of engagement with collections is to organize them; 36% of the subjects did this weekly, 57% did it monthly, and the last 7% reported they never did it at all. The last group corresponds, for example, to Abrams, "no-filers" [1]. The frequency of engaging in collection building/referring was again greater than that of collection organizing in a statistically significant manner [$F(2,26)=3.45$, $p<0.002$]. This shows that people refer to their collections as much as they build the collections, but they rarely organize their collections.

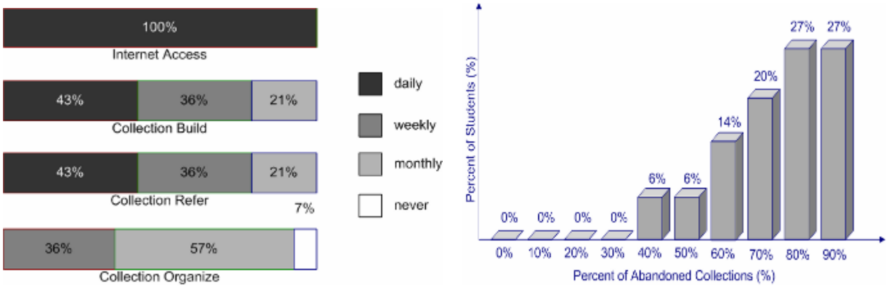


Fig. 1. Left - participants' internet access and collection building/referring/organizing frequency; Right - rate at which participants' collections are unutilized and abandoned

Collection Sharing

The study data shows that participants share their collections with other people, and also across several computers. 85% of participants said that they have their own blogs or personal web sites and publish some of their collections to share with others. These published collections may in turn function as source materials for others' collection building processes.

As mentioned above, one participant (P4) keeps URLs in email in order to access them from different computers. All participants said that they use several computers in different places. Participants use portable devices to carry their digital media materials or store them in network accessible spaces in order to share among different computers and as well as with others.

Breakdowns in Collection Practice

We investigated discrepancies between participants' expectations, and their experiences in practices of collecting. Our goal in identifying these breakdowns is to articulate user needs and design requirements. The most common breakdowns that participants experienced during the present study arose during their practices of referring, organizing, and finding things in their collections (P15: *"I initially made URL collections using bookmarks without any folder structure and renaming. Later, I had trouble finding a specific URL in it, so I deleted all my bookmarks and made folders with renaming. After this experience, I became more cautions about adding and renaming URLs to the collection."*). They said that they initially didn't have trouble finding elements in collections they built, but as time elapsed after collection building, it became more difficult to remember what is in the collections, and where. Recall, a problem of limited human attention, becomes a problem (P12: *"I had really important data in my collections, but I cannot find it! Could you make a program for me?"*). As the set of collections they own grows larger, it becomes difficult to remember all of them. Even though they sometimes don't have any clue of where the elements are, they said that they start browsing their collections first rather than searching. When they don't find the elements in the expected location, they use a search tool (P13: *"I seldom organize my collection very well, so I went through all folders one by one sequentially trying to find a certain file. Sometimes, I forgot what I saved, so I searched the web instead of the collections, and saved the same thing again."*). However, they may not even remember what to search for.

As mentioned above, 27% of participants said that collection building is somewhat useless because most parts of their collections are not utilized, and thus abandoned. Participants were asked what percent of their digital collections remain unutilized. At least 40% of the participants' collections are abandoned (See Figure 1 Right); 27% of participants said that 90% of their collections are abandoned; another 27% of participants indicate that 80% of their collections are abandoned; for 20% of participants 70% of collections are abandoned; 14% of participants have a 60% abandonment rate; 6% of participants have 50% abandoned collections; another 6% have 40% abandoned collections. Nonetheless, participants continue to engage in collecting (P4: *"Even though I am not using most of my collections and I sometimes think what I've built is useless, I keep building collections."*).

The participants initially build their collections with the intention of using them later. However, most collected material is not utilized because of trouble remembering and finding what has been collected. They lack effective means for referring to their collections. Collections are abandoned not because the information and media they contain are useless, but because of breakdowns in utilization practice.

Reasons for Collection Building

Participants were asked why they still build collections even though they do not utilize most parts of them. Like P14 (*"Wow, I realize that I am not using most parts of my collections, around 90%"*), they are often unaware that they are not utilizing most of what they collect. However, all participants still build collections from some sense that they will need the collected information elements later (P6: *"I want to save time on searching when I need a document in the future. That is my main reason for continuing to build collections."*). They collect media files to enjoy and also to share with others. Participants collect information that seems meaningful, useful and needed. They collect media that seems fun, unique, and consonant with their personal tastes. They make collections not for the definite promise of later utility, but from some intuitive sense of meaning and value.

4.2 Using Semantics to Represent Collections

Through the study, we observed that participants create semantic structures to organize their collections using any available affordances. They build their own structures for meaningfully representing their collections for usage later.

Developing Informal Metadata Schemas

All participants said that they make hierarchical directory structures to organize and manage their collections. They make folders based on contents, dates, semantic identifiers related to tasks or activities, or other categories that are somehow significant to them. Participants said that folder structures are created and changed because collections are added and deleted continuously.

Participants said that they rename files and file folders using metadata such as date, location, title, or author in order to help find them later. Renaming is important for search also. They seek to remember which words they used to rename files, in order to reuse them later when they browse and search their collections. Several participants mentioned strategies other than renaming for keep tracking of collected material. For example, they create index files inside of folders so that they can know what they contain (P6: *"Inside file folders, I make a 'readme' file to look at it later. This will help me to remember what the collection is about. In the individual file, I rename the file, and in addition to that, I put an explanation about the content in the first line."*).

We identify participants' practices such as renaming elements and creating hierarchical folder structures for representing important and large collections as the development of *informal metadata schemas*. They found ways to develop informal metadata schemas even in the absence of tools that support extensible field creation. They used the single accessible field afforded by existing tools that is the file or link name, to store the metadata. This practice was mostly spontaneous, occurring without an ontological plan. It was conducted informally and incrementally, as a series of situated actions [20]. This is an example of incremental formalism [18].

4.3 Suggestions

Participants were asked what new functionalities would be helpful in tools for collection building and utilization. Categories were not specified. Participants could mention whatever was on their minds. Participants' suggestions addressed areas such as collection utilization statistics display, filing assistance, and collection privacy support. They wanted help in renaming their collection materials in order to make the structure consistent, to make it easier to find materials later. They also asked for cues such as a 'visited count,' which shows how many times the owner read the file, in their collection representations and search and browsing environments to support finding specific materials. They liked the way desktop search is moving to assist collection utilization, however, they wanted their private files to be processed differently (P13: *"I have a big paper collection, but it is hard to find the paper I need when I need it using search tools supported in Windows. I tried Google desktop search, and it is pretty good, but one time I was a little embarrassed because a file that I wanted to keep private was retrieved as a search results when I was with my friend"*).

5 Discussion

Study participants invest substantial personal effort and resources into processes of building and utilizing collections. Their persistence in collecting in spite of breakdowns conveys the sense that they need to keep collecting to support a range of activities that span personal and work-related parts of their lives. In this section we examine participants' engagement with collections and the needs they express, and extrapolate from these, while considering human cognitive facilities and emerging technological capabilities. The result is to derive implications and ideas for designers of systems that support collection building and utilization.

The data shows that participants' breakdowns were centered in processes of collection utilization. They had trouble finding specific elements in their collections, and even though they built collections of elements that were useful, most of them are not utilized in the relevant context because of limited human attention and memory. They forget what to look for and where. Abandoned collections consume disk space, and more importantly, human attention during browsing, which is people's first choice for how to refer to collections.

We propose prescriptions to address breakdowns discovered in this study. Since the discovered breakdowns generally involve limitations of human understanding of collections, the prescriptions involve making better use of individual cognitive resources, sharing collections, and the definition of collection semantics. The first prescription addresses breakdowns that involve forgetting what has been collected, by using representations for collections that better cue human memory. The next proposed solution is based on ambient displays that use peripheral attention and changes over time for individual and collaborative interaction with collection visualizations. Other user needs that result from analysis of the breakdowns involve distributed tools for collection sharing, and the automatic generation of metadata schemas.

We can take steps to help people track of their collected information, by making better utilization of human memory capabilities. It is a well-accepted principle of cognitive science that in the working memory system, the visuospatial buffer, which store mental images, and the rehearsal loop used for text are complementary subsystems [2]. Thus, dual coding strategies that represent the elements stored in a collection with images as well as text will improve memory utilization [2][8], and contribute to helping people find elements while browsing. Thus, we can provide users with tools that support them in developing and generating visual index representations of their collections, which integrate images and text. These representations will be easier to remember, promote recognition, and facilitate the formation of mental models [10]. Since collection representations function as visual communication, either from a user to her/himself or between users, visual design principles must be applied during processes of collection organization..

Developing representations during collection-building and explicit organization activities is one solution. But people don't have sufficient attention to always work on representing their collections. Another prescription develops peripheral ambient visualizations that gradually display elements from collections over time. Ambient visualizations use time as a dimension in collection visualization. They can represent personal and group collections, engaging human attention without requiring it. Ambient visualizations can be deployed on a dedicated display, or as a screensaver. The set of collections that get visualized can be specified explicitly by users, and/or by an agent that uses clues, such as recency of access. For example, a large display in a collaborative environment such as a research lab or departmental work area can visualize collected materials that represent information relevant to current projects and research. This method can jog memories and promote serendipity, to facilitate individual and collaborative utilization of meaningful, useful and important elements in collections. Affordances that enable privacy will be required.

Additionally, we have seen that sharing with others is an important motivation for peoples' collecting practices. People utilize and collect information on multiple computers and devices in different locations. This can cause access problems, when the person is in one place, and the needed information is somewhere else. One initiative that addresses this is 'del.icio.us', which supports URL collection sharing [17]. del.ico.us enables users to tag URLs while collecting. It shows the metadata that others have used, and enables social browsing through these relationships. We believe this is a start for sharing collections and their semantics. New collection tools need to consider people's social and distributed collection-sharing intentions and enable collecting actual objects as well as references, while considering accessibility and privacy. Deeper semantic structures than single tags will also add value. These functionalities need to be integrated with editing, saving, browsing, and searching in order to best use limited human attention.

Users who are organizing collections by building informal metadata schemas need more powerful semantic structures. Easy to use extensible metadata systems will address this need. New collection tools need to use human attention effectively by supporting people's processes of semantic schema development in context, using content analysis, text pattern recognition, and image processing techniques. They can apply and extend collaborative filtering techniques for making suggestions about which metadata tags fit what is being saved [17][9]. Feature-based clustering and content analysis techniques can be applied to facilitate the semantic organization of collections by grouping similar information elements and building referential links. Users need to be able to override as well as accept the resulting suggestions. As part of this process, agents can track mutually relevant information elements scattered across the computer and the network, and inform the user about related information elements in diverse collection substructures using similarity measures.

6 Conclusion

Our study participants display tenacity in their involvement in processes of collecting. They explicitly express the intention and need to be involved in ongoing practices of collecting. They collect digital media materials involved in a broad range of activities, spanning personal and work relationships, which make up their everyday experiences. Their collection artifacts directly signify, relate to, and support these activities. Thus, collections and the process of collecting, itself, play important roles in how people create meaning in their lives.

Participants engage in collection building and utilizing activities regularly, even though it is not mandatory, and even though problems arise in the user experience. They keep collecting in spite of breakdowns. Better representations can help support these processes, by making better use of human attention. Tools for collecting need to be based in a sense of supporting individual and collaborative processes of meaning creation, while maximizing utilization of cognitive resources.

References

1. Abrams, D., Baecker, R., Chignell, M., Information archiving with bookmarks: personal Web space construction and organization, Proc. SIGCHI, April 18-23, 1998, p.41-48.
2. Baddeley, A.D., Is working memory working?, Quarterly Journal of Exp Psych, 44A, 1-31, 1992.
3. Bälter, O., Strategies for Organizing Email, Proc. of HCI on People and Computers XII, p.21-38, January 1997

4. Barreau, D., Context as a factor in personal information management systems, JASIS, 46(5):327-339, June 1995.
5. Barreau, D., Nardi, B., Finding and reminding: file organization from the desktop. ACM SIGCHI Bulletin 27, 3 (1995), 39-43.
6. Billsus, D., Hilbert, D., Maynes-Aminzade, D., Improving Proactive Information Systems, Proc. IUI 2005, January 9, 2005, p. 159-166
7. Boardman, R., Sasse, M. A., "Stuff goes into the computer and doesn't come out": a cross-tool study of personal information management, Proc. SIGCHI 2004, p. 583-590.
8. Carney, R.M., Levin, J.R., Pictorial Illustrations Still Improve Students' Learning From Text, Educational Psychology Review, Vol. 14, No. 1, March 2002.
9. Davis, M., King, S., Good, N., Sarvas, R., From context to content: leveraging context to infer media metadata, ACM Multimedia 2004, pp. 188-195.
10. Glenberg, A.M., Langston, W.E., Comprehension of illustrated text: Pictures help to build mental models, Journal of Memory & Language, 31(2):129-151, April 1992.
11. Hawkey, K., Inkpen, K. M., Privacy gradients: exploring ways to manage incidental information during co-located collaboration, Proc. CHI 2005, April, 2005, p. 1431-1434.
12. Jones, W., Dumais, S., Bruce, H., Once found, what then?: a study of "keeping" behaviors in personal use of Web information. Proc. ASIST 2002, November 18-21, 2002, 391-402.
13. Malone, T., How do people organize their desks?: Implications for the design of office information systems, TOIS, 1(1):99-112, Jan. 1983
14. Pew Internet & American Life Project, Internet: The Mainstreaming of Online Life, 2005, http://www.pewinternet.org/pdfs/Internet_Status_2005.pdf
15. Pew Internet & American Life Project, The Internet Goes to College, 2002, http://www.pewinternet.org/pdfs/PIP_College_Report.pdf
16. Pew Internet & American Life Project, Content Creation Online,http://www.pewinternet.org/pdfs/PIP_Content_Creation_Report.pdf
17. Schachter, J., del.icio.us, http://del.icio.us
18. Shipman, F., Marshall, C., Formality Considered Harmful: Experiences, Emerging Themes, and Directions on the Use of Formal Representations in Interactive Systems, Proc. CSCW 1999, 333-352.
19. Simon, H., Computers, Communications and the Public Interest, Martin Greenberger, ed., The Johns Hopkins Press, 1971, 40-41.
20. Suchman, L., Plans and Situated Actions, New York: Cambridge University Press, 1987.
21. Whittaker, S., Sidner, C., Email overload: exploring personal information management of email, Proc. SIGCHI, April 13-18, 1996, p.276-283.
22. Winograd, T., Flores, F., Understanding Computers and Cognition, Addison-Wesley, 1986

An Exploratory Factor Analytic Approach to Understand Design Features for Academic Learning Environments

Shu-Shing Lee, Yin-Leng Theng, Dion Hoe-Lian Goh,
and Schubert Shou-Boon Foo

Division of Information Studies
School of Communication and Information
Nanyang Technological University
Singapore 637718
{ps7918592b, tyltheng, ashlgoh, assfoo}@ntu.edu.sg

Abstract. Subjective relevance (SR) is defined as usefulness of documents for tasks. This paper enhances objective relevance and tackles its limitations by conducting a quantitative study to understand students' perceptions of features for supporting evaluations of subjective relevance of documents. Data was analyzed by factor analysis to identify groups of features that supported students' document evaluations during IR interaction stages and provide design guidelines for an IR interface supporting students' document evaluations. Findings suggested an implied order of importance amongst groups of features for each interaction stage. The paper concludes by discussing groups of features, its implied order of importance, and support for information seeking activities to provide design implications for IR interfaces supporting SR.

Keywords: Subjective relevance, exploratory factor analysis, interface design.

1 Introduction

Information retrieval (IR) systems are traditionally developed using the "best match" principle assuming that users can specify their needs in queries [3]. It retrieves documents "matching closely" to the query and regards these documents as relevant. Here, relevance is computed objectively using a similarity measure between query terms and terms in documents without considering users' needs and tasks [24].

This paper enhances objective relevance and addresses its limitations by taking a quantitative, subjective relevance (SR) approach. The SR concept provides suitable theoretical underpinnings for our approach as it focuses on document's relevance for users' needs [12]. This paper builds on an initial study [15] where features supporting users' evaluations of subjective relevance of documents were elicited. Here, we aim to understand university students' perceptions for elicited features. Specifically, we use factor analysis to investigate groups of features and their implied order of importance to provide design guidelines for IR interfaces supporting SR.

Our approach may show designers how users' perceptions of importance of features may be elicited and how factor analysis may be used to imply order of importance for features so that better decisions are made to design IR interfaces supporting

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 315 – 328, 2006.

users' relevance evaluations of documents. Similarly, our work applies to digital libraries by supporting designers determine design features for IR interfaces so that users are guided to find documents based on their needs.

2 Related Work

Different approaches have attempted to enhance objective relevance by developing user-centered IR systems. One method adopts an algorithmic approach to support techniques like collaborative browsing and collaborative filtering in IR systems. Collaborative browsing aims to understand how users interact with other users to facilitate browsing processes and retrieve relevant documents. An example application is *Let's Browse* [16]. Collaborative filtering helps users retrieve relevant documents by recommending documents based on users' behaviors and behaviors of similar users. Example applications are *Fab* [1] and *GroupLens* [23].

In the digital library domain, researchers have tried to design user-centered systems that helped users retrieve relevant documents. One such work is the *Digital Work Environment* library [18] which points users of a university digital library to relevant documents based on their user categories and tasks. Another example uses a participatory design approach through techniques like observations and low-tech prototyping to develop a user-centered children's digital library called *SearchKids* [9].

Another research area looks at user-centered criteria and dimensions affecting relevance judgments, such as, [2] and [19]. These works may allow IR designers to provide appropriate information that helps users find documents for tasks.

3 Theoretical Framework

Our approach differs from those highlighted in Section 2. Firstly, we focus on the location stage in the information life cycle [11], where we use SR to elicit features supporting users' relevance evaluations of documents. Secondly, we conducted a quantitative study identifying users' perceptions of elicited features. Factor analysis was used to discover groups of features for IR interaction stages and their implied order of importance amongst groups to provide design guidelines for IR interfaces supporting users' evaluations of relevance of documents for academic research.

This paper builds on our first study [15]. SR [6], information seeking in electronic environments [17], and a model of user interaction [21] were used to provide rationale for the first study. In that study, the SR concept was used to elicit features. SR was defined as usefulness of an information object for users' tasks [4]. SR also referred to different intellectual interpretations that a user conducted to interpret if an information object was useful [4]. The four SR types were [6]:

- Topical relevance: This relevance is achieved if the topic covered by the assessed information object corresponds to the topic in user's information need.
- Pertinence relevance: This relevance is measured based on a relation between user's knowledge state and retrieved information objects as interpreted by the user.
- Situational relevance: This relevance is determined based on whether the user can use retrieved information objects to address a particular task.
- Motivational relevance: This relevance is assessed based on whether the user can use retrieved information objects in ways that are accepted by the community.

The first study also investigated how stages in Marchionini's [17] model of information seeking were mapped to phases in Norman's [21] model of user interaction. The mapping aimed to illustrate how users might interact with an IR system to complete tasks. Our mapping showed that Marchionini's [17] model was similar to Norman's [21] model in terms of three stages (see Figure 1). It was implicitly inferred that Norman's [21] stages of task completion could be implied in each stage as each stage involved completing a task, such as, query formulation.

In the first study, subjects completed a task using exemplary IR systems. The task informed subjects to think about what features supported their relevance evaluation of documents. Subjects brainstormed SR features for IR interfaces. Elicited features were analyzed using SR types, stages in information seeking and phases in the model of user interaction to understand how students' used features during IR interactions. Features not coded to SR types were removed. Details of this study are found in [15].

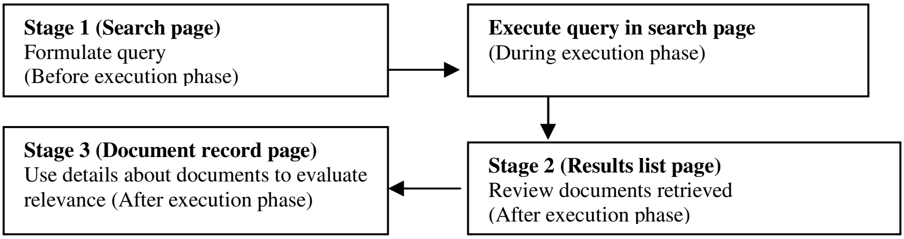

Stage 1 (Search page)
Formulate query
(Before execution phase)
Execute query in search page
(During execution phase)
Stage 3 (Document record page)
Use details about documents to evaluate relevance (After execution phase)
Stage 2 (Results list page)
Review documents retrieved
(After execution phase)

Fig. 1. Stages of Users' Interactions in IR Systems

4 A Study

Using digital libraries as examples of IR systems, we designed a survey form and conducted a study based on SR features from the first study. In an ideal situation, various methods, such as, reviewing IR systems and asking large groups of users could be used to get features for the survey. However these methods could yield many features and made decisions on what features to be included in the survey difficult.

The study was exploratory and aimed to gather students' perceptions of features elicited in the first study. Specifically, the study investigated students' perceptions of features as they imagined completing a task in a digital library. Data gathered were analyzed using exploratory factor analysis (EFA) as EFA removed redundant features and identified relationships so that groups describing most of the original data were discovered [14; 20]. Thus, groups of features supporting students' IR interaction stages could be identified to provide guidelines for designing IR interfaces supporting SR. Reasons for conducting the quantitative study was because a qualitative study could be expensive and time-consuming as there was a need to interview subjects, videotape and transcribe interviews. Moreover, the qualitative study might gather rich data with many relationships that made it difficult to remove redundant features and data gathered might not be generalisable to larger populations.

4.1 Designing the Survey Form

The designed survey form consisted of three parts:

- Part 1 provided a brief overview of the study.
- Part 2 included a glossary of difficult terms to help participants rate SR features.
- Part 3 consisted of two sections. Section A contained a list of 50 SR feature questions. A five-point Likert scale (very important; important; neutral; not very important; not important) was used to rate each SR feature. Our previous work [15] indicated that SR judgments were related to users' tasks and IR interactions. Hence, a task scenario and stages that students might experience were highlighted at the start of Section A. The IR interaction stages were: S1) formulate and execute query in the search page; S2) review documents in results list; and S3) view details in the document record page to support evaluation of documents. Participants considered the task and stages as they rated SR features. This approach was in line with Carroll's [5] scenario-based design. Section B contained demographic questions.

The form was pilot-tested with 2 self-reported information seeking experts and 2 novices. Their feedback indicated that questions might be organized by IR interaction stages. Analyses done in the first study [15] were used to re-organize questions.

4.2 Methodology

The survey form was handed out during 6 Master's level and 8 Undergraduate level classes. Participants rated their perceptions of importance of SR features based on a given scenario of use. 565 responses were received of which 465 were valid. A valid response was defined as a form that had all 50 SR feature questions answered.

Profiles of Participants
48.4% of students were males and 51.6% of students were females. Ages ranged from 18-49 years old and 65% were less than 23 years old. The high percentage of students younger than 23 years old was because most of them were undergraduates.

Data Analysis Method
EFA was conducted according to organization of questions into the 3 interaction stages. EFA was conducted using Principle Components Analysis with varimax rotation and a 0.4 factor loading. This factor loading was suitable for EFA [20].

Three heuristics were used to extract the number of factors for each analysis. In the first heuristic, factors were extracted above the "elbow" of the scree plot [14; 20]. The second heuristic extracted as many factors that had eigenvalues greater than 1 [14; 20]. The third heuristic was to compare eigenvalues from a dummy dataset with eigenvalues from the real dataset, and factors in the real dataset that had eigenvalues higher than those in the dummy dataset were retained [14]. These heuristics provided a range of factors to explore to derive the most meaningful factor solution. The most meaningful factor structure was selected using these criteria [7]: 1) the factor structure accounted for at least 50% of the variance amongst features included in the structure; 2) each factor had at least 3 features; 3) no or few cross factor loadings; and 4) factors must be meaningful. Reliability of each factor was checked using Cronbach's coefficient alpha [8]. A threshold value of 0.6 was selected [22]. If a factor had an alpha value below 0.6, items in the factor were removed and analysis was repeated.

It is emphasized that the final factor solution for each interaction stage was decided based on the criteria for most meaningful factor structure and we did not aim for each factor to account for more than 50% of the variance amongst features in the solution.

5 Findings and Discussion

Factors for stage 1 are described in detail. Due to limited space, findings for stages 2 and 3 are shown in tables and described briefly. We discuss the implied order of importance for factors in each stage and its implications towards interface design. Findings are also discussed in terms of how features support information seeking activities stated in Ellis' [10] behavioral model of information seeking.

5.1 Findings for Stage 1 (Search Page)

We started with a comprehensive set of 17 SR features for the search page. EFA reduced it to 14 features and loaded them to 3 factors. The factors accounted for 54.543% of the total variance (that is, the dispersion of data) in the 14 features. The features were coded to pertinence relevance in the first study [15], thus, factor names attempted to reflect this fact. SR features here were coded to pertinence relevance because success of determining pertinence relevance depends, to a certain extent, on the ability of users to formulate queries. In turn, users' ability to formulate queries is dependent on their knowledge of a topic or perceptions of information need [6].

Table 1 shows factor loadings for stage 1. Factors are labeled as S1_F1 to S1_F3 to indicate that it supported stage 1 and its respective factor number in this stage. Tables 2 and 3 are constructed similarly. Factors for stage 1 are described in detail below.

- *Factor S1_F1: Search Options for Query Formulation and Pertinence Relevance*

Features in Factor S1_F1 (see Table 1, column S1_F1) indicated search options that guided students formulate queries, especially for those who could not articulate their needs. Alpha value for this factor was 0.852.

Table 1. Factor Loadings of SR Features for Stage 1

SR features	Factor loadings		
	S1_F1	S1_F2	S1_F3
1. Search in journal title field	0.834		
2. Search in abstract field	0.799		
3. Search in author field	0.791		
4. Search in document full text	0.757		
5. Provide search tutorials and examples		0.695	
6. Provide advanced search mode		0.607	
7. Provide basic search mode		0.600	
8. Provide "clear query" button		0.563	
9. Provide search history		0.526	
10. Basic search considers query as a phrase if no Boolean operators are specified		0.494	
11. Method of entering and executing queries should be simple like search engines			0.720
12. Provide search entry boxes			0.664
13. Search in keywords field			0.431
14. Search in title field			0.404

- *Factor S1_F2: Additional Features for Query Formulation and Pertinence Relevance*

Factor S1_F2 described additional features supported query formulation in the search page. Example features were: provide basic and advanced search modes (see Table 1, column S1_F2 for all features). This factor's alpha value was 0.644.

- *Factor S1_F3: Basic Features for Query Formulation and Pertinence Relevance*

This factor included basic features that let students specify their queries, like, provide search entry boxes. Search options here supported query formulation for students who knew their information need, such as, keywords describing contents and titles of documents (see Table 1, column S1_F3 for features). The alpha value was 0.669.

5.2 Discussion for Stage 1 (Search Page)

Principles of EFA indicated that the first factor extracted would account for the highest percentage of total variance in all variables analyzed and subsequent factors would account for as much of the remaining variance as possible that was not accounted by the preceding factor [14]. Thus, the order in which factors were extracted and the percentage of total variance in all features analyzed were used to imply the order of importance for factors in each stage [13]. This rationale for implying order of importance was used to discuss findings for all stages.

- *Most Important SR Features for Stage 1*

Factor S1_F1 contained the most important SR features for stage 1 as it accounted for the highest amount of total variance in the 14 features analyzed for this stage (34.142%). This factor indicated different search options for the search page (see Table 1). Thus, students might have found search options to be most important as it showed the types of information that could be searched. Search options in Factor S1_F1 differed from those in Factor S1_3 (see Table 1, rows 13-14). This was because search options in Factor S1_F1 were more comprehensive and allowed students to search for documents using different means, such as, by author, abstract, or full text whereas search options in Factor S1_F3 seemed to support query formulation for students who knew the titles and keywords of documents they needed.

- *Second Most Important SR Features for Stage 1*

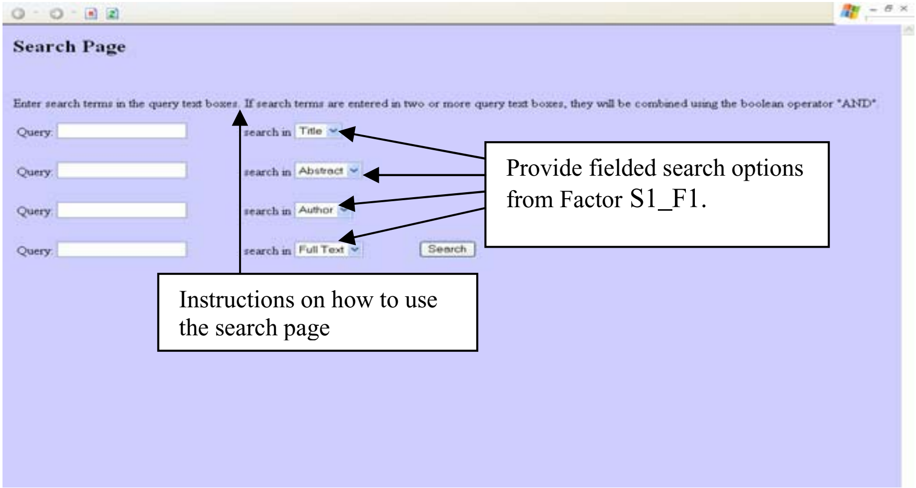
Features in Factors S1_F2 (see Table 1) were the second most important SR features as it was ranked second for percentage of total variance in the 14 features analyzed (11.293%). Thus, it was inferred that besides providing search options, students also wanted other features to support query formulation. For example, if different search modes were designed, students could select a search mode depending on their needs.

- *Third Most Important SR Features for Stage 1*

Features in Factor S1_F3 (see Table 1) were ranked third for the amount of total variance in the 14 features analyzed in stage 1 (9.109%). Reason could be because students felt that the feature, "provide search entry boxes", was redundant as search pages should have text boxes for users to enter queries. Factor S1_F3 was similar to Factor S1_F1 as search options were available in both factors. However, search options in Factor S1_F3 might not be as important as those from Factor S1_F1 as students might not know keywords or titles of relevant works. Thus, search options in Factor S1_F1 would provide more access points for students to search for documents.

Analyses of SR features for stage 1 yielded three factors ranked in implied order of importance. Hence, depending on students' needs and design resources, different groups of SR features might be designed in the search page. For example, if resources were limited, then the most important SR features in Factor S1_F1 could be designed. However, if comprehensive support for query formulation was needed then all three factors of SR features could be designed to provide basic and advanced search pages.

Features highlighted in factors for stage 1 seemed to support the information seeking activities of starting, browsing and monitoring. Features here might support starting as students could have initial references recommended by their teachers and they might formulate queries to find out if these documents were available in the system. Alternatively, students could already have a clear understanding of their need and were actively browsing (that is, semi-directed / semi-structured searching) to look for relevant documents or they could search the system to monitor developments within interested areas. Figure 2 shows the designed search page with most important SR features. Search option with highest factor loading was designed on the top and the one with the lowest factor loading was designed at the bottom.



Search Page

Enter search terms in the query text boxes. If search terms are entered in two or more query text boxes, they will be combined using the boolean operator "AND"

Query: search in Title

Query: search in Abstract

Query: search in Author

Query: search in Full Text

Provide fielded search options from Factor S1_F1.

Instructions on how to use the search page

Fig. 2. Search Page with Most Important SR Features

5.3 Findings for Stage 2 (Results List Page)

A comprehensive list of 21 SR features for stage 2 was packed to 5 factors. The factors accounted for 52.567% of the total variance in all 21 features. Factors are labeled as S2_F1 to S2_F5, factor loadings and alpha values are described in Table 2.

Factor S2_F1 was labeled "*point students to documents supporting topical, situational and motivational relevance*" as features (see Table 2, rows 1-5) were coded to these SR types and indicated different ways of pointing students to other documents. Features in Factor S2_F2 (see Table 2, rows 6-10) could help students find suitable contents and document types for their needs. Moreover, features were coded to topical, situational and motivational relevance in the first study [15]. Hence, this factor was named "*features for evaluating contents for topical, situational and motivational relevance*". Features for Factor S2_3 (see Table 2, rows 11-13) were coded to topical and situational relevance in the first study [15] so this factor was named "*alternate ways of presenting results list to support topical and situational relevance*". Factor S2_F4 was labeled "*extra information to evaluate documents for topical, situational and motivational relevance*" as features (see Table 2, rows 14-17) were coded to topical, situational and motivational relevance in the first study [15]. These features provided additional information about retrieved documents and its source to facilitate document evaluations. Features for Factor S2_F5 (see Table 2, rows 18-21) included those that were commonly available in results list and they were coded to topical relevance in the first study [15]. Hence, this factor was named "*common features available in results list page to support topical relevance*".

5.4 Discussion for Stage 2 (Results List Page)

- *Most Important SR Features for Stage 2*

Factor S2_F1 (see Table 2) were inferred as the most important SR features for stage 2 as it had the highest percentage of total variance in all features analyzed (26.060%). The survey form asked students to rate features with the assumption that the results list included a list of retrieved documents. Hence, it was inferred features in Factor S2_F1 could be built on top of retrieved documents in the results list page.

- *Second Most Important SR Features for Stage 2*

Features in Factor S2_F2 (see Table 2) focused on allowing students evaluate appropriate contents and document types for their needs. This factor was inferred as second most important because it was ranked second in terms of total variance in all features analyzed for stage 2 (7.911%).

- *Third Most Important SR features for Stage 2*

Factor S2_F3 (see Table 2) focused on providing novel ways of presenting results list and providing explanations of how documents were ranked. Features here might indicate that students were willing to try new ways of presenting documents in results list to determine if these methods were effective. Features in this factor were inferred as third most important because its percentage of total variance in all features analyzed was ranked third amongst factors extracted for stage 2 (6.912%).

- *Fourth Most Important SR features for Stage 2*

Factor S2_F4 focused on features that provided additional information to help students evaluate documents for their needs. Thus, if students could not get sufficient information, they might turn to features in Factor S2_F4 to get more information to support their document evaluations. Features here were implied as the fourth most important for stage 2 as its percentage of total variance in all features analyzed (5.916%) was ranked fourth amongst the five factors for this stage.

- *Fifth Most Important SR features for Stage 2*

Features in Factor S2_F5 (see Table 2) were inferred as fifth most important for this stage as its percentage of total variance in all features analyzed was ranked fifth (5.767%). Reason might be because students rated features based on their assumptions of common features in results lists. Hence, features here were redundant as they matched students' perspectives.

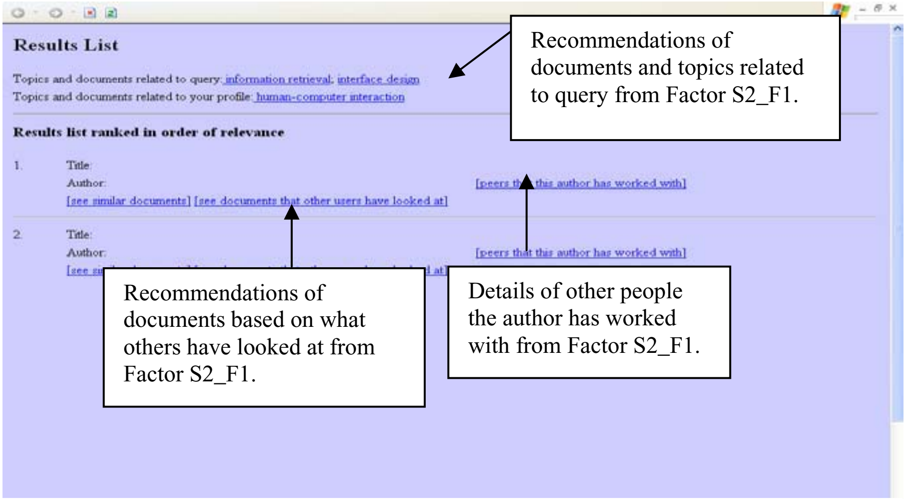
Table 2. Factor Loadings of SR Features for Stage 2

SR features	Factor loadings				
Factor S2_F1: Point students to documents supporting topical, situational and motivational relevance (Alpha value: 0.738)					
1. Recommend related documents and topics based on query	0.796				
2. Recommend related documents for each document retrieved	0.781				
3. Provide details of other people the author had worked with	0.603				
4. Recommend documents based on what others have looked at	0.461				
5. Recommend related documents based on user's profile and searching behavior	0.453				
Factor S2_F2: Features for evaluating contents for topical, situational and motivational relevance (Alpha value: 0.697)					
6. Provide an abstract for each document retrieved in results list		0.732			
7. Allow users to preview abstract before downloading full text		0.723			
8. Highlight search terms for each document in results list		0.676			
9. Provide an option so users can choose to display a paragraph or a few lines in which search terms appear in full text		0.502			
10. Categorize documents retrieved based on types of documents like journals, conference proceedings, etc.		0.447			
Factor S2_F3: Alternate ways of presenting results list to support topical and situational relevance (Alpha value: 0.643)					
11. Rank documents in results list in terms of how many times it has been used by others			0.720		
12. Provide explanation of how documents are ranked			0.713		
13. Present results list in pictorial format			0.491		
Factor S2_F4: Extra information to evaluate documents for topical, situational and motivational relevance (Alpha value: 0.614)					
14. Provide link that shows general information about document's source				0.631	
15. Provide link to document source's table of contents				0.615	
16. Provide subject categories for each document retrieved				0.610	
17. Provide selected references cited for each document retrieved				0.610	
Factor S2_F5: Common features available in results list page to support topical relevance (Alpha value: 0.617)					
18. Rank retrieved documents in results list in order of relevance					0.716
19. Display results list					0.660
20. Rank and provide relevance percentage for documents retrieved in results list					0.608
21. Allow searching within documents retrieved in results list					0.506

The factors seemed to include features that were exclusive to their respective factors except for an overlap amongst features in Factors S2_F3 and S2_F5. The overlapping occurred as features in both factors related to ranking of documents retrieved. However, there were slight differences. The feature in Factor S2_F3 (see Table 2, row 11) focused on ranking documents retrieved based on frequency of use whereas features in Factor S2_F5 (see Table 2, rows 18 and 20) focused on ranking documents in order of relevance and relevance percentage.

An order of importance was implied amongst factors for stage 2. Thus, features in different factors could be implemented as groups. Students might activate clusters and incrementally add features to the interface as pop-up boxes and pull-down menus

Features highlighted in factors for stage 2 seemed to support the information seeking activities of chaining and differentiating. Students might perform backward chaining by following references cited in documents to gain access to other documents. Backward chaining might be supported by the feature, "provide selected references cited for each document". Forward chaining was also supported by features in factors for stage 2 which involved providing links to other possible relevant documents through recommendation methods, such as, by users' profiles, and related topics. Most features in factors for stage 2 aimed to provide information to help students differentiate if a retrieved document was worth evaluating in more detail in the document record page. Examples of such features were: provide abstract, and categorize documents based on document type. Figure 3 illustrates the designed results list page incorporating most important features for stage 2 (Factor S2_F1). Features were built on top of a ranked list of retrieved documents.



Results List

Topics and documents related to query: information retrieval; interface design
Topics and documents related to your profile: human-computer interaction

Results list ranked in order of relevance

1. Title:
Author: [peers th this author has worked with]
[see similar documents] [see documents that other users have looked at]

2. Title:
Author: [peers that this author has worked with]
[see si d at]

Recommendations of documents and topics related to query from Factor S2_F1.

Recommendations of documents based on what others have looked at from Factor S2_F1.

Details of other people the author has worked with from Factor S2_F1.

Fig. 3. Results List Page with Most Important SR Features

5.5 Findings for Stage 3 (Document Record Page)

Twelve comprehensive features were loaded to 3 factors. Factor loadings, factor names and alpha values for stage 3 are shown in Table 3. The factors accounted for 58.959% of the total variance in the 12 features analyzed.

Factor S3_F1 was named "*seek others' help to evaluate documents for pertinence and motivational relevance*" as features identified (see Table 3, rows 1-4) were coded to pertinence and motivational relevance in the first study [15]. Features here seemed to allow students discuss relevance with authors and other users. Features in Factor S3_F2 (see Table 3, rows 5-9) were coded to situational relevance in our first study [15] and facilitated management of full text. Thus, this factor was labeled "*features that support access and management of full text for situational relevance*". Factor S3_F3 (see Table 3, rows 10-12) provided full text and highlighted search terms so students could evaluate relevance of highlighted text in relation to contents.

5.6 Discussion for Stage 3 (Document Record Page)

- *Most Important SR Features for Stage 3*

Features in Factor S3_F1 were inferred as the most important features as its percentage of total variance in all features analyzed was the highest (33.822%). Students rated features based on an understanding that the document record page provided detailed information, such as, title, author and publisher. Hence, it was inferred that students were keen to discuss with others to find relevant documents and features here could be built on top of detailed information in document record page.

Table 3. Factor Loadings of SR Features for Stage 3

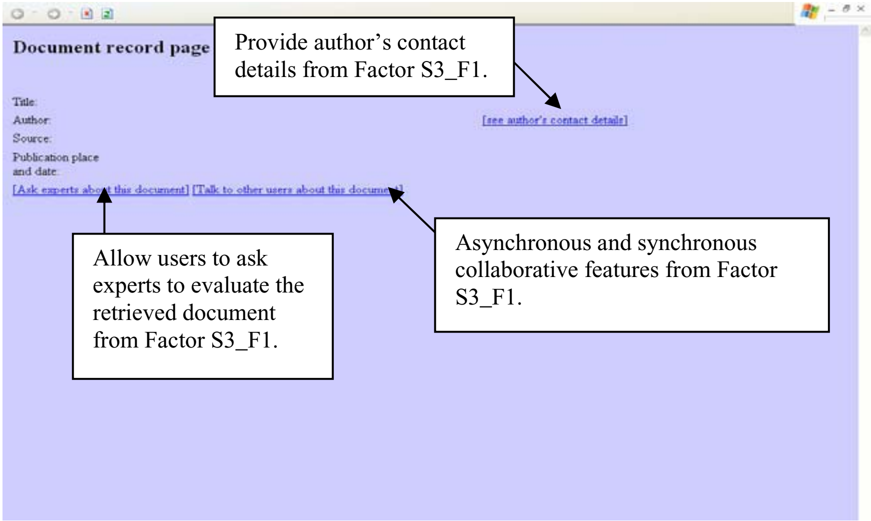
SR features	Factor loadings		
Factor S3_F1: Seek others' help to evaluate documents for pertinence and motivational relevance (Alpha value: 0.795)			
1. Provide asynchronous collaborative features	0.896		
2. Provide synchronous collaborative features	0.869		
3. Provide author's contact details	0.653		
4. Allow users to ask experts to evaluate documents retrieved	0.652		
Factor S3_F2: Features that support access and management of full text for situational relevance (Alpha value: 0.761)			
5. Allow full text to be saved using its title as the default file name		0.823	
6. Allow full text to be saved in a compressed version		0.794	
7. Print full text without "highlighted / bolded" search terms		0.628	
8. Provide "reader" software in the document record page		0.623	
9. Specify on what pages in full text do search terms appear and provide link to the page		0.459	
Factor S3_F3: Highlight portions in full text and point users to other documents for situational relevance (Alpha value: 0.657)			
10. Highlight search terms in full text			0.830
11. Provide links to full text of documents cited in the current document			0.676
12. Allow users to download full text in PDF format			0.676

- *Second Most Important SR Features for Stage 3*

Factor S3_F2 focused on providing features that facilitated access and management of full texts. Hence, it was inferred that students wanted easy access and management of full texts so that they would extract relevant content for tasks. Features here were deduced as the second most important features as its percentage of total variance in all features analyzed (15.233%) was ranked second amongst factors for stage 3.

- *Third Most Important SR Features for Stage 3*

Features in Factor S3_F3 were specified as third most important as its percentage of total variance in all features analyzed (9.904%) was ranked third amongst factors for



Document record page

Title:
Author:
Source:
Publication place and date:

[see author's contact details]

[Ask experts about this document] [Talk to other users about this document]

Provide author's contact details from Factor S3_F1.

Allow users to ask experts to evaluate the retrieved document from Factor S3_F1.

Asynchronous and synchronous collaborative features from Factor S3_F1.

Fig. 4. Document Record Page with Most Important SR Features

this stage. Reasons could be: 1) students wanted to read full text to extract information; and 2) students might find full text of cited documents to be relevant.

The three factors extracted for stage 3 seemed to indicate that three important groups of features could be designed. Features in these groups seemed unique and there were no overlaps. Thus, depending of design requirements different groups of important features could be designed. Features indicated in factors for Stage 3 seemed to support the information seeking activities of differentiating and extracting. This was because the document record page provided detailed information so that students could differentiate if the retrieved document was useful. Moreover, the document record page also provided access to full text so that students could extract contents.

Figure 4 shows the designed document record page with most important SR features. As students rated features based on an understanding that the document record page provided detailed information about the document, like, title, author and publisher, features in Factor S3_F1 were built on top of such information.

6 Conclusion and On-Going Work

Our approach differs from approaches addressing collaborative browsing and filtering, user-centered design approaches and user-defined criteria for relevance judgments highlighted in Section 2. Firstly, our approach used SR as a theoretical basis to elicit features supporting document evaluations. We also used stages of IR interaction to understand how students might use features to complete tasks in IR systems. Secondly, we investigated students' perceptions for elicited features using EFA. The contributions of our work are:

- EFA extracted groups of SR features to support each stage of students' IR interactions. Although all groups of features were important to form the factor solutions to support students' document evaluations during IR interactions, there seemed to be an implied order of importance amongst groups. Thus, depending on requirements, different groups of features could be designed in IR interfaces.
- The groupings seemed to indicate clusters of SR features that could be implemented collectively. Student might activate different clusters and features could be added to the interface in the form of pop-up boxes and pull-down menus.

Findings presented are preliminary and have limitations. The study gathered students' perceptions of importance of SR features without actually using the system. Students might have different understandings of SR features and this could be problematic when students did not have prior experience using such features. Hence, future work may focus on verifying and evaluating our findings in a qualitative study where users could comment on importance of SR features in actual context of use. Findings presented are exploratory and applied specifically to students who participated in the study. Future work might use EFA to discover groups of SR features supporting IR interactions for other students in different task scenarios so that insights could be gathered on the needs of larger student populations for IR interfaces supporting SR. The translation of factors into interface design is also another area that needs to be looked into in future.

References

1. Balabanovic, M. and Shoham, Y. (1997). Fab: Content-based, collaborative recommendation. *Communications of the ACM, 40* (3), 66-72.
2. Barry, C. L. (1994). User-defined relevance criteria: an exploratory study. *Journal of the American Society for Information Science, 45* (3), 149-159.
3. Belkin, N. J., Oddy, R. N., and Brooks, H.(1982). ASK for information retrieval: Part I. background and theory. *The Journal of Documentation, 38* (2), 61-71.
4. Borlund, P. and Ingwersen, P. (1998). Measures for relative relevance and ranked half-life: Performance indicators for interactive IR. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* ACM Press, 324-331.
5. Carroll, J. M. (2000). *Making use: Scenario-based design of human-computer interactions.* California, USA: The MIT Press.
6. Cosijin, E., and Ingwersen, P. (2000). Dimensions of relevance. *Information Processing and Management 63,*533-550.
7. Costello, A. B. and Osborne, J. W. (2005). Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation: A Peer-reviewed Electronic Journal, 10* (7), http://pareonline.net/pdf/v10n7.pdf.
8. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika 16*, 297-334.
9. Druin, A. et al. (2001). Designing a digital library for young children: An intergenerational partnership. *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries,* ACM Press, 398-401.
10. Ellis, D. (1989). A behavioural approach to information retrieval system design. *Journal of Documentation, 45* (3), 171-212.
11. Fischer, G., Henninger. S. and Redmiles, D. (1991). Cognitive tools for locating and comprehending software objects for reuse, *Proceedings of 13th International Conference on Software Engineering*, IEEE Computer Society Press, 318-328.
12. Ingwersen, P. and Borlund, P. (1996). Information transfer viewed as interactive cognitive processes. In Ingwersen, P. and Pors, N. O. (Eds.). *Information Science: Integration in Perspective.* Royal School of Librarianship, Denmark, 219-232.
13. Kim, Jae-On and Mueller, C. W. (1978). *Factor analysis: statistical methods and practical issues.* California: Sage Publications, Inc.
14. Lattin, J., Carroll, J. D., and Green, P. E. (2003). *Analyzing multivariate data*. Nelson, Canada: Brooks/Cole.
15. Lee, S. S., Theng, Y. L., Goh, D. H. L, and Foo, S. S. B. (2005). Subjective relevance: implications on interface design for information retrieval systems. In Fox, E., Neuhold, E. J., Pimrumpai, P, and Wuwongse, V. (Eds.), *The 8th International Conference on Asian Digital Libraries, ICADL, 2005. Digital libraries: implementing strategies and sharing experiences* (pp. 424-434). Germany, Berlin: Springer-Verlag.
16. Lieberman, H. (1995).An agent for web browsing. *Proc. International Conference on Artificial Intelligence,* 924-929.
17. Marchionini, G. (1995). *Information seeking in electronic environments.* Cambridge, UK: Cambridge University Press.
18. Meyyapan, N., Chowdhury, G. G. and Foo, S. (2001). Use of a digital work environment prototype to create a user-centered university library. *Journal of Information Science, 27* (4), 249-264.

19. Mizzaro, S. (1998). How many relevances in information retrieval?. *Interacting with Computers, 10*, 303-320.
20. Netemeyer, R. G., Bearden, W. O., and Sharma, S. (2003). *Scaling procedures: issues and applications.* California, USA: Sage Publications.
21. Norman, D. A. (1998). *The psychology of everyday things*. New York: Basic Books.
22. Nunnally, J.C. (1978).Psychometric Theory (2nd ed.). New York: MacGraw-Hill.
23. Resnick, P., Iacovou, N., Mitesh, S., Bergstron, P., and Riedl, J. (1994). GroupLens: an open architecture for collaborative filtering of Netnews. *Proc. ACM Conference on Computer Supported Cooperative Work,* ACM Press, 175-186.
24. Tang, R. and Soloman, P. (1998). Toward an understanding of the dynamics of relevance judgment: An analysis of one person's search behavior. *Information Process and Management 34,*237-256.

Representing Contextualized Information in the NSDL

Carl Lagoze[1], Dean Krafft[1], Tim Cornwell[1], Dean Eckstrom[1],
Susan Jesuroga[2], and Chris Wilper[1]

[1] Computing and Information Science, Cornell University, Ithaca, NY 14850 USA
{lagoze, dean, cornwell, eckstrom, cwilper}@cs.cornell.edu
http://www.cis.cornell.edu
[2] UCAR-NSDL, PO Box 3000, Boulder, CO 80307 USA
jesuroga@ucar.edu

Abstract. The NSDL (National Science Digital Library) is funded by the National Science Foundation to advance science and math education. The initial product was a metadata-based digital library providing search and access to distributed resources. Our recent work recognizes the importance of context – relations, metadata, annotations – for the pedagogical value of a digital library. This new architecture uses Fedora, a tool for representing complex content, data, metadata, web-based services, and semantic relationships, as the basis of an information network overlay (INO). The INO provides an extensible knowledge base for an expanding suite of digital library services.

1 Introduction

Libraries, traditional and digital, are by nature information rich environments - the organization, selection, and preservation of information are their raison d'etre. In pursuit of this purpose, libraries have focused on two areas: building a collection of all the *resources* that meet the library's selection criteria, and building a catalog of *metadata* that facilitates organization and discovery of those resources.

This is the approach that the NSDL (National Science Digital Library) Project took over its first three years of existence, when it focused mainly on the location and development of resources appropriate for Science, Technology, Engineering, and Mathematics education, and the creation of quality metadata about those resources. This focus was reflected in the technical infrastructure that harvested metadata from distributed providers, processed and stored that metadata, and made it available to digital library services such as search and preservation.

The value of an excellent collection of resources as a basis for library quality is undeniable. And, even after years of advances in automatic indexing, metadata remains important for a class of resources and applications. However, our three years of effort in the NSDL have revealed that collection building and metadata aggregation are necessary but not sufficient activities for building an information-rich digital library. In particular, our experience has led to two conclusions. First, the technical and organizational infrastructure to support harvesting, aggregation, and refinement of metadata is surprisingly human-intensive and expensive [15]. Second, in a world of increasingly powerful and ubiquitous search engines, digital libraries must distinguish themselves by providing more than simple search and access [16]. This is particularly

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 329 – 340, 2006.

true in educationally-focused digital libraries where research shows the importance of interaction with information rather than simple intake.

Based on these conclusions, we have redirected our efforts over the past year towards building a technical infrastructure that supports a more refined definition of information richness. This definition includes, of course, collection size and integrity, and it accommodates the relevance of structured metadata. But it adds the notion of building *information context* around digital library resources. Our goal is to create a knowledge environment that supports aggregation of multiple types of structured and unstructured information related to library resources, the instantiation of multiple relationships among digital library resources, and participation of users in the creation of this context. We are creating an infrastructure that captures the wisdom of users [32], adding information from their usage patterns and collective experience to the formal resources and structured metadata we already collect.

Our technical infrastructure is based on the notion of an *information network overlay* [16] – a directed, typed graph that combines local and distributed information resources, web services, and their semantic relationships. We have implemented this infrastructure using Fedora [17], an architecture for representing complex objects and their relationships.

In this paper we describe the motivations for this architecture, present the information model that underlies it, and provide results from our year of implementation. We note for the reader that this is still a work in progress. The results we provide in this paper relate to the implementation and scaling issues in creating a rich information model. As our work progresses, we will report in future papers on the effectiveness of this architecture from the perspective of the user and evaluate whether it really does enable a richer and more useful digital library.

The organization of this paper is as follows. Section 0 describes related work and situates this work in the context of other digital library efforts. Section 3 summarizes the importance of information contextualization for educational digital libraries. Section 4 provides a brief background on the NSDL and establishes the application context in which this work occurs. Section 0 describes the information model of the information network overlay. Section 6 provides the results of our implementation experience. Finally, section 0 concludes the paper.

2 Related Work

The work described in this paper builds on a number of earlier and ongoing research and implementation projects that investigate the role of user annotations in information environments, the importance of inter-resource relationships, and the integration of web services with digital content. We believe that our work is distinguished from these other projects in two ways. First, it combines traditional digital library notions of resources and structured metadata with service-oriented architecture and semantic web technology, thereby representing the rich relationships among a variety of structured, unstructured, and semi-structured information. Second, it implements this rich information environment at relatively large scale (millions of resources), exercising a number of state-of-the-art technologies beyond their previous deployments.

Perhaps the most closely related work is the body of research on information annotation. Catherine Marshall has written extensively on this subject [20] in the digital library and hypertext context. A number of systems have been developed that implement annotation in digital libraries. For example, Roscheisen, Mogensen, and Winograd created a system call ComMenter [31] that allowed sharing of unstructured comments about on-line resources. The multi-valent document work at Berkeley provides the interface and infrastructure for arbitrary markup and annotation of digital documents, and storage and sharing of that markup [34]. The semantic web community has also examined annotation, with the Annotea project [13] being the most notable example.

The importance of annotation capabilities for education and scholarly digital libraries has been noted by many researchers including Wolfe [35]. The ScholOnto project [24] created a system for the publication and discussion/annotation of scholarly papers, arguing for the importance of informal information along-side established resources. Constantopoulus, et al. [8] examine the semantics of annotations in the SCHOLNET project, a EU-funded project to build a rich digital library environment supporting scholarship. Within the NSDL effort, there have been a number of projects that support annotations, most notably DLESE (Digital Library for Earth System Education) [1].

Annotations and their association with primary resources are one class of the variety of relationships that can be established among digital content. Ever since Vannevar Bush invented hypertext [5], researchers have been examining tools for interlinking information. Faaborg and Lagoze [11] examined the notion of semantic browsing whereby users could establish personalized and sharable semantic relationships among existing web pages. Huynh, et al. [12] have recently done similar work in the Simile project.

There is also related work on resource linking specifically for pedagogic purposes within the educational research community. Unmil, et al. [33] describe Walden's Paths, a project that allows teachers to establish meta-structure over the web graph for creation of lesson plans and other learning materials. Recker, et al. have created another system called Instructional Architect [28], that similarly allows integration of on-line resources by teachers into educational units.

Finally, an important component of the work described here is the integration of content and web services. In many ways our digital library "philosophy" resembles that of the Web 2.0 philosophy [25]. Key components of this are the collection and integration of unique data, the participation of users in that data collection and formulation process, and the availability of the data environment as a web service that can be leveraged by value-add providers. Chad and Miller [6] extend Web 2.0 to something they call Library 2.0. We hope that our work demonstrates many of the principles they describe, notably the notion that Library 2.0 encourages a "culture of participation" and provides the interface to its accumulated information for innovative "mash-ups" that exploit library information in innovative ways.

3 The Need for Context and Reuse

Research shows that education-focused digital libraries (and digital libraries in general) need to support the full life cycle of information [19]. Reeves wrote "The real power of media and technology to improve education may only be realized when students actively use them as cognitive tools rather than simply perceive and interact with them as tutors or repositories of information." [30].

One requirement that appears frequently in the learning technology literature is the reuse of resources for the creation of new learning objects. This involves integrating and relating existing resources into a new learning context. A learning context has many dimensions including social and cultural factors; the learner's educational system; and the learner's abilities, preferences and prior knowledge [21].

Most digital libraries, including the NSDL, currently rely on forms of metadata to describe learning objects and enable discovery. Metadata standards abstract properties of learning objects, and abstraction can lead to instances where learning context is ignored or reduced to single dimensions [26]. Metadata is often focused on the technical aspects of description and cataloging, not on capturing the actual context of instructional use. Recker and Wiley write "a learning object is part of a complex web of social relations and values regarding learning and practice. We thus question whether such contextual and fluid notions can be represented and bundled up within one, unchanging metadata record." [29]

McCalla also argues that there is no way of guaranteeing that metadata captures the breadth and depth of content domains. He writes that, ideally, learning objects need to reflect "appropriateness" to address the differences between learners' needs. [22] In addition, questions remain as to whether these logical representations (e.g. metadata and vocabularies) created primarily for use by computer systems will make the most intuitive sense for learners [7].

Several approaches have been suggested to help supply the rich context for learning object creation and reuse. These include capturing opinions about learning objects and descriptions of how they are used [26]; recording the community of users from which the learning object is derived [29]; collecting teacher-created linkages to state education standards [28]; tracking and using student-generated search keywords [2]; and providing access to comments or reviews by other faculty and students [23].

We see that in order to provide an educationally-focused digital library, the information infrastructure must support flexible integration of information, ranging from highly structured metadata to unstructured comments and observations. It needs to nr dynamic, expanding both in the manner that a standard library collection expands, but also based on the collective experience and input of the user community.

4 A Suite of Contextualized NSDL Services

We are creating the infrastructure to meet notions of information richness outlined in the previous section. This work follows more than three years of work by the NSDL Core Infrastructure (CI) team, and has been described in a number of other papers

[14, 15]. Stated very briefly, this earlier work used OAI-PMH to populate a metadata repository (MR). This metadata was indexed by a CI-managed search service, which was accessible by users through a central portal at http://nsdl.org.

Our goal is to move beyond the search and access capabilities provided by the MR. The creation of the NSDL Data Repository (NDR), built on the architecture described in the next section, provides a platform for a number of exciting new NSDL applications focused directly on increasing user participation in the library. In addition to creating specific new capabilities for NSDL users, these applications all create context around resources that aids in discovery, selection and use. Four specific applications that exploit the infrastructure described in this paper are currently in various phases of development, testing, and deployment.

Expert Voices (EV) is a collaborative blogging system that fully integrates the resources and capabilities of the NDR. It allows subject matter experts to create real-time entries on critical STEM issues, while weaving into their presentation direct references to NSDL resources. These blog entries automatically become both resources in the NSDL library and annotations on all the referenced resources. EV supports Question/Answer discussions, resource recommendations and annotations, the provision of structured metadata about existing resources, and establishing relationships among existing resources in the NSDL, as well as between blog entries and resources.

On Ramp is a system for the distributed creation, editing, and dissemination of content from multiple users and groups in a variety of formats. Disseminations range from publications like the NSDL Annual Report to educational workshop materials to online presentations like the Homepage Highlights exhibit at NSDL.org's homepage. Resources created and released in OnRamp can become NDR content resources, and NDR resources and other content can be directly incorporated into On Ramp publications, creating new context and relationships within the NDR.

Instructional Architect, described by Recker [27], "... enables users (primarily teachers) to discover, select, and design instruction (e.g., lesson plans, study aids, homework) using online learning resources. ". Currently, IA supports searching the NSDL for resources and incorporating direct references to those resources into an IA project. The IA team is currently working with the NDR group to support both publication of IA projects as new NSDL resources and the direct capture the web of relationships created by an IA project in the NDR.

The *Content Alignment Tool* (CAT), currently in development by a team led by Anne Diekema and Elizabeth Liddy of Syracuse University, uses machine learning techniques to support the alignment of NSDL resources to state and national educational standards [10]. Initially (2Q2006), users will be able to use the tool to suggest appropriate educational standards for any resource they are viewing. Later versions of the system will allow experts and other users to provide feedback, incorporated into the NDR, on the appropriateness of the assignments. This tool, and the overall incorporation of educational standards relationships into the NDR, will allow NSDL users to search and browse the NSDL by "standards", starting either from a standard or from any relevant resource.

5 Design and Information Model

To provide the foundation for this rich array of user-visible services, we have implemented the NSDL Data Repository (NDR). The NDR implements all features of the pre-existing MR such as metadata harvesting, storage, and dissemination. However, it moves from the restrictive metadata-centric focus of the MR to a resource-centric model, which allows representation of rich relationships and context among digital library resources.

The NDR implements a data abstraction that we call an information network overlay (INO). Like other overlay networks [3] the INO instantiates a layer over another network, in this case the web graph.

Specifically, an INO is a directed graph. Nodes are identified via URIs and are packages of multiple streams of data. This data stream composition corresponds to compound object formats such as METS [18] and DIDL [4], allowing the creation of compound digital objects with multiple representations. The component data streams may be contained data or they may be surrogates (via URLs) to web-accessible content. This allows nodes to aggregate local and distributed content, for example the reuse of multiple primary resources into new learning objects. Web services may be associated with information units and their components, allowing service-mediated disseminations of the data aggregated in a digital object. This advances the reuse paradigm beyond simple aggregation, allowing, for example, a set of resources written in English to be refactored into a Spanish learning object though mediation by a translation service. Edges represent ontologically-typed relationships among the digital objects. The relationship ontology is extensible in the manner of OWL-based ontologies [9]. This allows the NDR to represent the variety of application-based relations described earlier such as collection membership, aggregation via reuse into a learning object, and correlation with one or more state standards. Nodes (digital objects) are polymorphic - they can have multiple types in the data model, where typing means the set of operations that can be performed on the digital object. In the digital library environment, this flexibility overcomes well-known dilemmas such as the data/metadata distinction, which conflicts with the reality that an individual object can be viewable through both of these type lenses.

The NDR is implemented within a Fedora repository. A complete description of Fedora is out-of-scope for this paper and the reader is directed to the up-to-date explanation at [17]. Each node in the INO corresponds to a Fedora digital object. Fedora provides all the functionality necessary for the INO including compound objects, aggregation of local and distributed content, web service linkages, and expression of semantic relationships. Fedora is implemented as a web service and includes fine-grained access control and a persistent storage layer.

Length constraints on this paper prohibit a full description of the information modeling in the NDR and the use of Fedora to accomplish this modeling. This modeling includes the design of Fedora digital objects to provide the different NDR object types – resources, agents, metadata, aggregations, and the like – and the relationships among these types for common use cases such as resource and metadata branding and resource annotation.

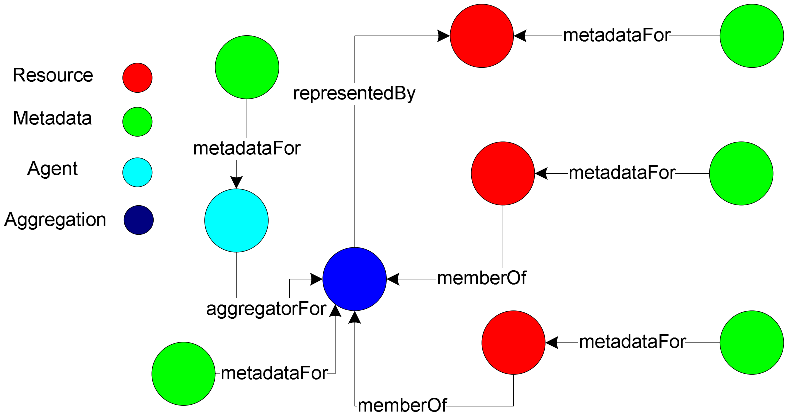


Fig. 1. Modeling an aggregation

However, an example shown in Fig. 1 demonstrates how the NDR represents aggregation. Examples of aggregations include conventional collection/item membership, but also aggregations with other semantics such as membership of individual resources in a compound learning object or alignment of set of resources with a state educational standard. Each node corresponds to a Fedora digital object, with the key at the left showing the type of the object. The labels on the arcs document the type of the relationship. As shown, "memberOf" arcs relate resources to one or more aggregations. Aggregations can have arbitrary semantics, with the semantics documented by the resource that is the object of the "representedBy" arc. For example, this resource may be a surrogate for a collection, or may represent a specific state standard. Lastly, the person or organization responsible for the aggregation is represented by the agent that is the source of the "aggregatorFor" arc.

6 Results from Implementation of the NSDL Data Repository

Over the past year we have been designing, implementing, and loading data into the NDR. The major implementation task was the creation and coding of an NDR-specific API for manipulation of information objects in the NDR data model – specific "types" of digital objects such as resources, metadata, agents, and the like and the required relationships among them. Note that this API is distinct from the SOAP and REST API in Fedora that provides access to low-level digital object operations. The NDR API consists of a set of higher level operations such as addResource, addMetadata, and setAggregationMembership. Each of these higher level operations is a composition of low-level Fedora primitive operations. For example, the logical NDR operation addResource, which adds a new resource to the NDR, translates to a set of low-level Fedora operations including creating the digital object that corresponds to the resource, configuring its datastreams so that they match our model for the resource "type", and establishing the relationships between that resource and its collection digital object and to the metadata digital objects that describe it.

We have implemented in Java an API layer that mediates all interaction with the NDR, by calling on the constituent set of low-level Fedora operations. In addition to providing a relatively easy-to-use interface for services accessing the NDR, the API performs the vital task of ensuring that constraints of the data model are enforced. For example, the data model mandates that no metadata digital object should exist that does not have one (and only one) "metadataFor" relationship to a resource digital object.

We have used this API to bootstrap the production NDR with data from the preexisting MR, thereby duplicating existing functionality in the new infrastructure. At the time of writing of this paper, this process is complete. The platform for our NDR production environment is a Dell 6850 server with dual 3Ghz Xeon processors, 32Gb of 400Mhz memory and 517Gb of SCSI RAID disk with 80MB/second sustained performance. This server is running 64-bit LINUX, for reasons outlined later. We note that the 2006 cost for this production server is about 22K USD.

The NDR has over 2.1 million digital objects – 882,000 of them matching metadata from the MR, 1.2 million of them representing NSDL resources, and several hundred representing other information objects – agents, services, etc., - in the NDR data model. The representation of the relationships among these objects (those defined by the NDR data model and those internal to the Fedora digital object representation) produces over 165 million RDF triples in the triple-store. We have found that ingest into the NDR takes about .7 seconds per object – making data load for this rich information environment a non-trivial task.

This bootstrapping process has been a learning process in scaling up semantically-rich information environments. In order to understand the results, it is necessary to distinguish three components: core Fedora, the triple-store it uses to represent and query inter-object relationships, and the Proai[1] component that supports OAI-PMH.

Core Fedora is a web service application built on top of a collection of file-system resident XML documents (one file for each digital object) and a relational database that caches fragments and transformations of those documents for performance. These XML documents are relatively small and stable, and at present we are using about 21 GBytes of disk space to store these files across 39,000 directories. We have not experienced any scaling problems nor do we foresee any with this core architecture. In fact, as we expected from our knowledge of the Fedora implementation, basic digital object access is not really dependent on the size of the Fedora repository. For example, our tests on dissemination performance show that requests for metadata formats that are stored literally in the NDR are about 69 ms. Requests for formats that are crosswalked from stored formats using an XSLT transform service take about 480 ms.

The more challenging aspect of our data loading and implementation work has involved the triple-store. Relationships among Fedora digital objects, and therefore among nodes in the NDR graph, are stored persistently as RDF/XML in a datastream in the digital object and are indexed as RDF triples in a triple-store, which provides query access to the relationship graph. In the case of the NDR, this provides query functionality such as "return all resources related to a state standard, a specific collection, or in an OAI set".

[1] http://www.fedora.info/download/2.1/userdocs/services/oaiprovider-service.html

Triple-store technology is relatively immature. Scaling it up to accomplish our initial data load has been especially challenging. As part of our implementation of the Fedora relationship architecture (known as the resource index), we experimented with scaling and performance of a number of tripe-store implementations. Our extensive tests comparing Sesame[2], Jena[3], and Kowari[4] are available online[5]. One particular target of our testing was the performance of complex queries that involve multiple graph node joins – these are the types of queries we issue to perform OAI-PMH List Records operations that select according to metadata format, set, and date range. We found that Jena would not scale over a few tens of thousands of triples with complex query times approaching 20 minutes for complex queries over .5 million triples. Sesame can be configured in both native storage mode or on top of mysql. We found that Sesame-mysql, like Jena, was unable to return large results sets, producing an out-of-memory error due to accumulating the entire result set in memory. Our remaining tests comparing Sesame native to Kowari showed that for a database of several million triples Kowari was faster by a factor of 2 for simple queries, and by a factor of over 9000 for complex queries.

Although the Kowari implementation proved capable under controlled tests of high performance and scalability, we encountered a number of hurdles along the path of our data load. The apparent reality is that neither Kowari nor any other triple-store has been pushed to this scale. Such scale revealed unpleasant and previously undiscovered bugs, such as a memory leak that took months of effort to verify and find[6]. Furthermore, we have found that the hardware requirements to run a large-scale semantic web application are non-trivial. Kowari uses memory mapped indexes, which are both disk and memory-intensive. Presently the Kowari-based resource index requires over 54 GB of virtual memory, which is significantly larger than the 5 GB addressable by standard 32-bit processors and operating systems (thus the configuration of our production server described earlier).

In order to understand our results on semantic queries to the NDR resource index (storing 165 million triples), it makes sense to divide these queries into two classes. The first class of queries is relatively simple, such as those issued by a user application seeking all resources correlated with a state standard or another accessing all members of a collection. We have found that query performance in this case is on the order of 25ms for the simplest examples (no transitive joins over the graph) to about 250 ms for examples with 2-3 joins. The second class of queries are those that populate the NDR OAI server, Proai, which is a part of the Fedora service framework. Proai is an advanced OAI server that supports any metadata format available through the Fedora repository via direct datastream transcription or service-mediated dissemination. It operates over a MySQL database that is populated via resource-index queries to Fedora (in batch after an initial load and incrementally over the lifespan of the Fedora repository). The resource-index queries to populate Proai are quite complex with semantics such as “list all Fedora disseminations representing OAI-records of a certain format, and get their associated properties and set membership information”.

[2] http://www.openrdf.org/

[3] http://jena.sourceforge.net/

[4] http://www.kowari.org/

[5] http://tripletest.sourceforge.net/

[6] http://prototypo.blogspot.com/2005/09/kowari-memory-leak-found-and-fixed.html

Such a query takes about 1 hour, when issued in batch over the fully loaded repository, and the combination of queries to pre-load the Proai database after the batch NDR load takes about 1-2 days. We note, however, that this load is only performed once on initial load of the NDR and that incremental updates, as information is added to the NDR, are much quicker.

Proai performance is quite impressive. Throughput on an OAI-PMH ListRecords request is about 900 records per/second, and we have been able to harvest all Dublin Core records from the NDR (to populate our search indexes) in about 3 hours.

Our results provide hardware guidelines for large Fedora implementations that use the resource index. We have found that they greatly benefit from a machine with large real memory, high-speed disks, and high-performance disk controllers. The Dual Xeon processors provide an excellent match for Fedora processing allowing uniform execution partitioning of core Fedora, the NDR API, Proai and MySql processing among the 4 hyper threaded CPU cores available. CPU clock rate is a minor performance factor compared with the overall memory and I/O performance of the chassis. As of this writing, machines with more than 32GB of memory remain rare. Within 18 months we anticipate that machines having 64GB will become commonly available.

7 Conclusions

We have described in this paper our initial work in implementing an advanced infrastructure to support an information-rich NSDL. This infrastructure supports the integration and reuse of local and distributed content, the integration of that content with web services, and the contextualization of that content within a semantic graph.

The work described in this paper has advanced the state-of-the-art in two areas. First, it involves the innovative use of Fedora to represent an information network overlay. This data structure combines local and distributed content management, service-oriented architecture, and semantic web technologies. At a time when digital libraries need to move beyond the search and access paradigm, the INO supports contextualized and participatory information environments. Second, this work pushes the envelope on scaling issues related to semantic web technologies. Although RDF and the semantic web have existed for over 8 years, large-scale implementations still need to be demonstrated. Our experience with scaling the NDR is instructive to a number of other projects looking to build on top of semantic web technologies.

The results in this paper demonstrate only the basic functionality of the NDR. The basic operations, however, are the building blocks for the applications described in Section 4. In future papers, we will describe our experience with these applications and the ability of the NDR to support them in a highly scaled manner.

Acknowledgements

We thank the entire NSDL CI team for their contributions to this work. The authors acknowledge the contributions of the entire Fedora team, especially Sandy Payette. The work described here is based upon work supported by the National Science Foundation under Grants No. 0227648, 0227656, and 0227888. Support for Fedora is provided by the Andrew W. Mellon Foundation.

References

1. Annotation Metadata Overview, http://www.dlese.org/Metadata/annotation/
2. Abbas, J., Norris, C. and Soloway, E., Middle School Children's Use of the ARTEMIS Digital Library. in *ACM/IEEE Joint Conference on Digital Libraries (JCDL '02)*, (Portland, OR, 2002), ACM Press, 98-105.
3. Andersen, D.G., Balakrishnan, H. and Kaashoek, M.F., Resilient Overlay Networks. in *18th ACM SOSP*, (Banff, Canada, 2001).
4. Bekaert, J., Hochstenbach, P. and Van de Sompel, H. Using MPEG-21 DIDL to Represent Complex Digital Objects in the Los Alamos National Laboratory Digital Library. *D-Lib Magazine*, *9* (11).
5. Bush, V.F. As We May Think *Atlantic Monthly*, 1945.
6. Chad, K. and Miller, P., Do Libraries Matter? The rise of Library 2.0, http://www.talis.com/downloads/white_papers/DoLibrariesMatter.pdf
7. Collis, B. and Strijker, A. Technology and Human Issues in Reusing Learning. *Journal of Interactive Media in Education*, *4* (Special Issue on the Educational Semantic Web).
8. Constantopoulos, P., Doerr, M., Theodoridou, M.*, et al.* On Information Organization in Annotation Systems. in Grieser, G. and Tanaka, Y. eds. *Intuitive Human Interface 2004, LNAI3359*, Springer-Verlag, Berlin, 2004, 189-200.
9. Dean, M., Connolly, D., van Harmelen, F.*, et al.* OWL Web Ontology Language 1.0 Reference. W3C Working Draft, 20020729.
10. Diekema, A. and Chen, J., Experimenting with the Automatic Assignment of Educational Standards to Digital Library Content. in *Joint Conference of Digital Libraries (JCDL)*, (Denver, 2005).
11. Faaborg, A. and Lagoze, C. Semantic Browsing. in *Lecture Notes in Computer Science*, Springer-Verlag, Trondheim, Norway, 2003, 70-81.
12. Huynh, D., Mazzocchi, S. and Karger, D., Piggy Bank: Experience the Semantic Web Inside Your Web Browser. in *International Semantic Web Conference (ISWC)*, (2005).
13. Kahan, J., Koivunen, M.-R., Prud'Hommeaux, E.*, et al.*, Annotea: An Open RDF Infrastructure for Shared Web Annotations. in *WWW10*, (Hong Kong, 2001).
14. Lagoze, C., Arms, W., Gan, S.*, et al.*, Core Services in the Architecture of the National Digital Library for Science Education (NSDL). in *Joint Conference on Digital Libraries*, (Portland, Oregon, 2002), ACM/IEEE.
15. Lagoze, C., Krafft, D., Cornwell, T.*, et al.*, Metadata aggregation and "automated digital libraries": A retrospective on the NSDL experience. in *Joint Conference on Digital Libraries*, (Chapel Hill, NC, 2006), ACM.
16. Lagoze, C., Krafft, D.B., Payette, S.*, et al.* What Is a Digital Library Anyway? Beyond Search and Access in the NSDL. *D-Lib Magazine*, *11* (11).
17. Lagoze, C., Payette, S., Shin, E.*, et al.* Fedora: An Architecture for Complex Objects and their Relationships. *International Journal of Digital Libraries*, *December 2005*.
18. Library of Congress, METS: An Overview & Tutorial, http://www.loc.gov/standards/mets/METSOverview.v2.html
19. Marshall, B., Zhang, Y., Chen, H.*, et al.*, Convergence of Knowledge Management and E-Learning: the GetSmart Experience. in *ACM/IEEE Joint Conference on Digital Libraries (JCDL '03)*, (Houston, TX, 2003), ACM Press, 135-146.
20. Marshall, C.C., Annotation: from paper books to the digital library. in *Digital Libraries '97*, (1997), ACM Press.
21. Martin, K. Learning in Context *Issues of Teaching and Learning*, 1998.

22. McCalla, G. The Ecological Approach to the Design of E-Learning Environments: Purpose-based Capture and Use of the Information about Learners. *Journal of Interactive Media in Education*, *7* (Special Issue on the Educational Semantic Web).
23. McMartin, F. and Terada, Y., Digital Library Services for Authors of Learning Materials. in *ACM/IEEE Joint Conference on Digital Libraries (JCDL '02)*, (Portland, OR, 2002), ACM Press, 117-118.
24. Motta, E., Shum, S.B. and Domingue, J. ScholOnto: an ontology-based digital library server for research documents and discourse. *International Journal on Digital Libraries*, *3* (3).
25. O'Reilly, T., What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software, http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html
26. Parrish, P. The Trouble with Learning Objects. *Educational Technology Research and Development*, *52* (1). 49-67.
27. Recker, M., Dorward, J., Dawson, D*., et al.* Teaching, Designing, and Sharing: A Context for Learning Objects. *Interdisciplinary Journal of Knowledge and Learning Objects*, *1*. 197-216.
28. Recker, M., Dorward, J. and Nelson, L.M. Discovery and Use of Online Learning Resources: Case Study Findings. *Educational Technology and Society*, *7* (2). 93-104.
29. Recker, M. and Walker, A. Collaboratively filtering learning objects. in Wiley, D.A. ed. *Designing Instruction with Learning Objects*, 2000.
30. Reeves, T.C., The Impact of Media and Technology in Schools: A Research Report prepared for The Bertelsmann Foundation, http://www.athensacademy.org/instruct/media_tech/reeves0.html
31. Roscheisen, M., Mogensen, C. and Winograd, T. Shared Web Annotations as a Platform for Third-Party Value-Added, Information Providers: Architecture, Protocols, and Usage Examples. Technical Report, CS-TR-97-1582,
32. Surowiecki, J. *The wisdom of crowds : why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. Doubleday :, New York, 2004.
33. Unmil, P.K., Francisco-Revilla, L., Furuta, R.K*., et al.*, Evolution of the Walden's Paths Authoring Tools. in *Webnet 2000*, (San Antonio, TX, 2000).
34. Wilensky, R. Digital library resources as a basis for collaborative work. *Journal of the American Society for Information Science*, *51* (3). 228-245.
35. Wolfe, J.L., Effects of Annotations on Student Readers and Writers. in *Fifth ACM International Conference on Digital Libraries*, (San Antonio, TX, 2000).

Towards a Digital Library for Language Learning

Shaoqun Wu and Ian H. Witten

Department of Computer Science
University of Waikato
Hamilton, New Zealand
{shaoqun, ihw}@cs.waikato.ac.nz

Abstract. Digital libraries have untapped potential for supporting language teaching and learning. Although the Internet at large is widely used for language education, it has critical disadvantages that can be overcome in a more controlled environment. This article describes a language learning digital library, and a new metadata set that characterizes linguistic features commonly taught in class as well as textual attributes used for selection of suitable exercise material. On the system is built a set of eight learning activities that together offer a classroom and self-study environment with a rich variety of interactive exercises, which are automatically generated from digital library content. The system has been evaluated by usability experts, language teachers, and students.

1 Introduction

The rise of computer-assisted language learning on the Internet has brought a new dimension to language classes. The Web offers learners a wealth of language material and gives students opportunities to learn in different ways. They can study by reading newspaper articles, listening to audio and viewing video clips; undertake online learning exercises; or join courses. Media such as email, chat and blogs enable them to communicate with other learners and with speakers of the target language all over the world. When preparing lessons, teachers benefit from the panoply of resources that the web provides. Automated tools can be used to build practice exercises and design lessons. Teachers construct language learning tasks based on the Internet because the language is real and the topics are contemporary, which motivates learners.

Despite all these advantages, the Internet has many drawbacks for language study. Although it offers innumerable language resources, learners and teachers alike face the challenge of discovering usable material. Search engines return an overwhelming amount of dross in response to any query, and locating suitable sources demands skill and judgment. When learners study on their own, it is hard for them to locate material that matches their language ability. Finally, students may accidentally encounter material with grossly unsuitable content.

Digital libraries, like traditional ones, can play a crucial role in education. Marchionini [1] identifies many advantages in using them for teaching and learning. As well as providing a safe and reliable educational environment, they have special advantages for language classes. Digital libraries are a great source of material that

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 341 – 352, 2006.

teachers can turn into meaningful language exercises. They offer vast quantities of authentic text. Learners experience language in realistic and genuine contexts, which prepares them for what they will encounter in the real world. Searching and browsing facilities can be tailored to the special needs of language learners. Teachers can integrate digital libraries into classes that help students locate appropriate material, giving them the tools to study independently. Interpersonal communication media can be incorporated to create a socially engaging learning environment.

This project has built a language learning digital library called LLDL based on the Greenstone digital library software [2]. The goal is to explore the potential of digital libraries in this field by addressing issues intrinsic to language learning. We developed a language learning metadata set (LLM) that characterizes linguistic features commonly taught in class. By using it in searching and browsing, teachers and learners can locate appropriate material.

Eight learning activities are implemented that utilize LLDL's search and retrieval facilities. Together they offer a classroom and self-study environment with a rich variety of interactive exercises. Four features distinguish them from existing systems:

- They are student-centered
- They provide a communicative learning environment
- They provide a multilingual interface
- Exercises are automatically generated from digital library content.

While the present implementation of LLDL is for learning English, it is designed to provide a multilingual interface. English and Chinese versions exist; new languages can easily be added. We close the paper with some remarks on extending the interface and the language taught to other European languages.

2 DLs in Language Learning

Digital libraries can serve many roles in language education. First, they provide linguistic resources. In the classroom, text, pictures, models, audio, and video are used as material for teaching. Edge [3] summarizes three kinds of language resource, published, authentic and teacher-produced, and digital libraries allow teachers to build collections of each kind. Culturally situated learning helps students interpret the target language and master skills in communication and behavior within the target culture [4]. Teachers can build collections that introduce the people, history, environment, art, literature, music. The material can be presented in diverse media—text, images, audio, video, and maps. Students can experience the culture without leaving the classroom.

Second, digital libraries can bring teachers and learners together. Forums, discussion boards, electronic journals and chat programs can be incorporated to create a community where teachers share their thoughts, tips and lesson plans; learners meet their peers and exchange ideas; and teachers organize collaborative task-based, content-based projects. This community is especially meaningful for language learning because it embeds learners in an authentic social environment, and also integrates the various skills of learning and use [5]. As Vygotsky [6] points out, true learning involves socialization, where students internalize language by collaborating on common activities and sharing the means of communicating information.

Third, digital libraries can provide students with activities, references and tools. Language activities include courses, practice exercises, and instructional programs. In traditional libraries students find reference works: dictionaries, thesauri, grammar tutorials, books of synonyms, antonyms and collocations, and so on. Equivalent resources in digital libraries can be used as the basis of stimulating educational games.

3 Language Learning Metadata

Metadata is a key component of any digital library. It is used to organize resources and locate material by searching and browsing. Metadata schemas developed specifically for education and training over the past few years have recently been formally standardized [7]. The two most prominent are LOM (Learning Object Metadata) and SCORM (Sharable Content Object Reference Model). LOM aims to specify the syntax and semantics of the attributes required to describe a learning object. It groups features into categories: general, life-cycle, meta-metadata, educational, technical, rights, annotation, and classification. SCORM aims to create flexible training options by ensuring that content is reusable, interoperable, durable, and accessible regardless of the delivery and management systems used. While LOM defines metadata for a learning object, SCORM references a set of technical specifications and guidelines designed to meet the needs of its developers, the US Department of Defense.

Neither of these standards proved particularly useful for our purpose. The aim of metadata is to help users find things. Although digital libraries make it easy to locate documents based on title, author, or content, they do not make it easy to find material for language lessons—such as texts written for a certain level of reading ability, or sentences that use the *present perfect* tense. To identify these users would have to sift through countless examples, most of which do not satisfy the search criteria.

The LLM metadata set is designed to help teachers and students locate material for particular learning activities. It has two levels: documents and sentences. All values are intended to be capable of being extracted automatically from full text. Some LLM metadata are extracted with the help of tools from the OpenNLP package, which provides the underlying framework for linguistic analysis of the documents by tagging all words with their part of speech and identifying units such as prepositional phrases.

3.1 Document Metadata

Readability metadata can help both teachers and students locate material at an appropriate level. We have adopted two widely used measures recommended by practicing teachers: Flesch Reading Ease and the Flesch-Kincaid Grade Level [8]. The former is normally used to assess adult materials, and calculates an index between 0 and 100 from the average number of words per sentence and the average number of syllables per word. The latter is widely used for upper elementary and secondary material and scores text on a US grade-school scale ranging from 1 to 12.

LLM incorporates both these scores as separate pieces of metadata, and in addition computes LOM *Difficulty* metadata by quantizing the Grade Level into five bands.

3.2 Sentence Metadata

Readability metadata is associated both with the document as a whole and with individual sentences. Three further types of metadata are associated with sentences: sentence metadata, syntactic metadata, and usage metadata.

LLDL splits every document into individual sentences using a simple heuristic involving terminating punctuation, the case of initial words, common abbreviations, and HTML tags. Whereas sentences used as examples in the classroom or language teaching books have been carefully targeted, prefabricated, and honed into clean and polished examples, sentences extracted automatically from authentic text are often untidy and incomplete; some have inordinately complex structures.

LLM addresses this by defining the following metadata for each sentence:

- Processed version
- Tagged version
- State: clean or dirty
- Type: simple or complex.

The first two are variants of the original extracted sentence, which usually contains HTML mark-up. The *Processed* version contains plain text: mark-up has been stripped. The *Tagged* version has been annotated with linguistic tags that reflect the syntactic category of each word. Part-of-speech metadata is used by the language learning digital library to generate exercises, as described in Section 5.

Some extracted sentences are messy. *State* metadata is used to indicate whether a sentence is *clean*, comprising alphabetic characters and punctuation only, or *dirty*, including other extraneous characters. The *Type* of a sentence is *simple* if it has just one clause and *complex* otherwise, where a clause is a group of words containing a main verb. Teachers normally use simple sentences to explain grammar rules where possible.

The extraction process first detects sentence boundaries and strips HTML, yielding *Processed* sentence metadata. If sentences contain any characters other than alphabetic ones, space, and punctuation, their *State* metadata is *Dirty*. Clean sentences are analyzed by the OpenNLP tagger and chunker to yield *Tagged* sentence metadata. These contain syntactic tags that reflect the categories of individual words and reveal the sentence structure, facilitating the extraction of language metadata. Simple and complex sentences are differentiated by the number of verb phrases (*VP*) they contain.

3.3 Syntactic Metadata

English grammar is relatively simple because it has fixed rules. On other hand, the number of rules is large and there are many exceptions. Based on recommendations from language teachers, we identified nine syntactic metadata elements that can be extracted automatically by natural language processing tools. While these do not cover all aspects of English grammar, they form the basis of a useful digital library.

The syntactic metadata elements are Adjective, Preposition, Possessive pronoun and determiner, Modal, Tense, Voice, Coordinating conjunction, Subordinate conjunction, That-clause and wh-clause. For each one a regular expression is defined—

for example, \\w+/JJ is the expression for *Adjective* metadata: it indicates a string that contains one or more word characters (\\w+) followed by */JJ*, the syntactic tag for adjective. *Tense* and *Voice* metadata are also extracted using tagged sentences. They comprise both the tense or voice and the verbs or verb groups that are so marked.

The extraction process for the remaining syntactic metadata is similar. Understanding the grammatical implications of the tags is the key to successful extraction. *Preposition* metadata is extracted by searching for prepositional phrases, tagged *PP*. *Subordinate conjunction* and *that-clause* metadata are extracted by seeking subordinating clauses tagged as *SBAR*. *Wh-clauses* are not indicated by a clause-level tag, and must be identified by combining phrase tags and *wh-word* tags.

3.4 Usage Metadata

LLM contains a single usage metadata element: Collocation. This is a group of two or more words that are commonly found together or in close proximity. For example, native speakers usually prefer the collocation *heavy rain* to the non-collocation *big rain*, or *totally convinced* to *absolutely convinced*. Lewis [9] points out that native speakers carry hundreds of thousands, possibly millions, of collocations in their heads ready to draw upon in order to produce fluent, accurate and meaningful language, and this presents great challenges to language learners.

We define collocations in terms of 9 two- and three-word syntactic patterns such as *adjective+noun*, *adverb+adjective*, and phrasal verbs in the form *verb+preposition*—for example, *make up* and *take off*. They are identified by looking for particular tags and matching them with the nine syntactic collocation patterns. Following common practice [10] we use the *t*-statistic to rank potential collocations. This uses the number of occurrences of words individually and in combination, and the total number of tokens in the corpus. Its accuracy depends on the size of the corpus: good collocations that occur just once do not receive high scores.

4 Searching the Digital Library

LLM metadata captures linguistic aspects of the documents in a digital library. It allows users to search and browse language learning materials. This section demonstrates the use of the extracted metadata in LLDL. In this project, we have built five demonstration collections for use in the activities described in the next section:

- Documents from the UN FAO *Better farming series*
- Children's short stories from *East of the web*
- News articles from the *BBC World Service*
- Sample articles from *Password*, a magazine for new English speakers
- Collection of plant and animal images downloaded from the Internet.

The first collection includes practical articles intentionally written in a simple style, but not targeted at children. The second contains material specifically for children. The third and fourth are made from material that is intended to be particularly suitable for second language learners. These four collections exhibit a wide variety of styles and difficulty levels.

LLDL uses standard Greenstone facilities [2] to present options for browsing and searching on entry to the library. When users browse, they can select *Titles*, *Difficulty*, and other metadata elements. Clicking *Titles* presents an alphabetical list of titles of the documents in the collection, broken down into alphabetic ranges; the full text of the documents is available by clicking beside the appropriate title. *Difficulty* also applies to documents, and allows the reader to browse titles in each of the five difficulty levels mentioned above.

The other browsing options refer to individual sentences: they are *Tense*, *Preposition*, *Clause*, *Difficulty* (which differs from the document-level *Difficulty* above because it refers to individual sentences), and *Type*. Sentences are the essential units in language communication. Students study vocabulary and learn grammars in order to construct sentences. Conversely, studying good sentences helps master word usage or grammar rules in context. LLDL allows readers to browse for particular grammatical constructions or identify particular parts of speech. For example, selecting *Preposition* shows the sentences of the collection, with the prepositions that each one contains listed in parentheses after it. The sentences are presented in alphabetic groups according to preposition: those under the *A–B* section of the hierarchy contain *about*, *at*, *above*, *as*, *between*, *before*, *by*, *beside*, … These sample sentences help students learn the usage of particular prepositions and study what words commonly appear before and after them—for example, *above all*, *ask about*.

Searching is more highly targeted than browsing. Users can perform an ordinary full-text search to locate documents that treat particular topics; the search results show the title and difficulty level of matching documents. Advanced search allows users to specify metadata as well as content. For example, one might search for particular full-text content but confine the search to documents that are *easy* (in terms of difficulty level). Or search for individual sentences rather than documents, whose type is *simple* (i.e., no compound sentences), or whose state is *clean* (i.e., no non-alphabetic characters). Users can combine these criteria in a search form to find *simple* and *clean* sentences from *easy* documents whose text contains specified words or phrases.

Users can also search for sentences that contain particular words. New learners are often confused about word usage—for example, distinguishing the different implications of *look*, *see* and *watch*. One way to help is to provide many authentic samples that show these words in context. LLDL can retrieve sentences that include a specified word or phrase, and are restricted by the above-mentioned sentence-level metadata. Students can also search for sentences that exhibit any of the grammatical constructs that are identified by metadata, for example passive voice sentences, modal sentences or sentences in a particular tense.

5 Language Learning Activities

LLDL facilitates the creation of language learning activities. To demonstrate this we have developed eight activities: *Image Guessing*, *Collocation Matching*, *Quiz*, *Scrambled Sentences*, *Collocation Identifying*, *Predicting Words*, *Fill-in-blanks*, and *Scrambled Documents*; unfortunately space permits a description of the first four activities only. They share the common components *login*, *chat*, *scoring* and *feedback*.

5.1 Common Components

Learners are not required to register, but must **login** by providing a user name and select a difficulty level (easy, medium or hard). This parameter is used to select sentences or documents for each activity, to determine which image collections are used to generate exercises, and to group students for activities in which they work in pairs. For these activities the system maintains a queue of users waiting at each level. When a student logs in, the queue is checked and they are either paired up with a waiting student at the same level, or queued to await a new opponent.

LLDL makes a **chat** facility available in all activities, in order to create an environment in which students can practice communication skills by discussing with peers, seeking help, and negotiating tasks. The chat panel resides either in the activity interface or a window that is launched by clicking a *Chat* button.

Each activity contains a **scoring** system intended to maintain a high level of motivation by encouraging students to compete with each other informally. Students can view the accumulated scores of all participants, sorted so that the high scorers appear at the top. Additional statistical information is provided such as the number of identified collocations in the *Collocation* activity or the number of predicted words in the *Predicting Words* activity. The implementation of the scoring mechanism varies from one activity to another, depending on whether students do the exercise individually, or collaborate in pairs, or compete in pairs.

Students are provided with **feedback** on whether the response is correct or incorrect, and in the latter case they are invited to try again, perhaps with a hint that leads to the correct response. In general, feedback is given item by item, at logical content breaks, at the end of the unit or session, or when requested by the student. Students also see their accumulated scores. Some activities provide an exercise-based summary that includes questions, correct answers, and answers by the student's partner.

Hints provide direct help without giving away the answer. They can be offered through text, pictures, audio or video clips, or by directing students to reference articles or relevant tutorials. Some exercises give hints that use text from the digital library. For example, the *Quiz* activity allows students to ask for other sentences containing the same words; *Collocation Matching* provides more surrounding text so that students can study the question in context.

5.2 The *Image Guessing* Exercise

In *Image Guessing*, the system pairs students according to their self-selected difficulty level. One plays the role of describer, while the other is the guesser. An image is chosen randomly from a digital library collection of images and shown to the describer alone; the guesser must identify that exact image. The describer describes the picture in words that are automatically used by the system as a query term, and also decides how many of the search results the guesser will see. The guesser does not see the description; the describer does not see the search results. The guesser and describer can communicate using the chat facility. The activity moves to the next image when the guesser successfully identifies the image, chooses the wrong one, or the timer expires. The students use the search and chat facility to identify as many images as possible in a given time. They can pass on a particular image, or switch roles.

The difficulty level is determined by the image collection, which teachers build for their student population. They select simple images—e.g. animal images or cartoons—for lower level students, and more complex ones—e.g. landscapes—for advanced ones. For searching, image collections use metadata provided by the teacher, which they tailor to the students' linguistic ability. The more specifically the metadata describes the images, the easier the game.

5.3 The Collocation Matching Exercise

Collocations are the key to language fluency and competence. Lewis [9] believes that fluency is based on the acquisition of a large store of fixed or semi-fixed prefabricated items. Hill [11] points out that students with good ideas often lose marks because they don't know the four or five most important collocations of a key word that is central to what they are writing about. Today, teachers spend more time helping students develop a large stock of collocations; less on grammar rules.

LLDL is particularly useful for learning collocations because it contains a large amount of genuine text and provides useful search facilities. In the *Collocation matching* activity, students compete in pairs to match parts of a collocation pattern. This is a traditional gap filling exercise in which one part of a collocation is removed and the students fill the gap with an appropriate word. For example, for *verb+preposition* collocations, verbs or prepositions are deleted. Students select the collocation type they want to practice on, and decide which component will be excised. The exercises use complete sentences retrieved from the library as question text.

Students are paired up and one is chosen to control the activity by selecting collocation types. The other one can see what is going on and negotiate using chat. Then complete sentences are presented one by one, with the target collocation colored blue and missing words replaced by a line. The students select the most appropriate word from four choices before the count-down timer expires. When the exercise is complete the pair view their performance in a summary window that shows the question text with collocations highlighted, and the students' answers and scores.

Exercises are generated from collocation metadata. Sentences at the appropriate difficulty level and collocation type are retrieved. The words that appear in the collocations are grouped according to their syntactic tags and used as choices for the exercise. For each sentence, four choices, including the correct one, are picked randomly.

5.4 The *Quiz* Exercise

Quizzes comprising a question and a few choices from which the correct answer must be selected are widely used language drills for learning grammar and vocabulary. Traditionally, teachers construct quizzes and use them for practice exercises, tests or exams. Our system offers a unique feature that makes quizzes far more motivational: students can create their own.

The teacher begins by defining a list of topics and perhaps creating a few initial quizzes. Students can select a topic and construct a new quiz by entering up to four words or phrases; limiting the maximum number of questions; choosing whether or not to stem the terms; and specifying sentence types—simple, complex or both.

Once the learner has defined a new quiz or selected an existing one, the system presents the questions. Each has two to five possible answers. When the student selects one, the system indicates its correctness and moves to the next question. Students can get help by initiating a digital library search for sentences that contain the correct word or words, without being told which one it is. When the quiz is finished a summary is shown of all questions, along with the correct answer and the student's incorrect ones. Students then re-take the questions on which they performed poorly.

This activity uses a simple quiz-generation mechanism that constructs questions and answers using words or phrases specified by students. For example, a question might be *What did you think ___ the film?* with possible answers *of*, *at*, *about*, and *over*. The question text comprises a single sentence retrieved from the digital library using words or phrases specified by the student. These are excised from the question text and used as the correct answer. Sentence retrieval employs full text search on the sentence text and metadata. For example, to construct questions on prepositions, teachers retrieve sentences by searching on *Preposition* metadata. To avoid students having to understand the metadata structure, they are only asked to provide the words or phrases of interest when creating new quizzes.

Stemming is a key parameter for quiz generation that significantly affects the number of available questions and choices. Without stemming, the question text for a *make* and *do* quiz would be restricted to sentences that contain *make* or *do*, and students would have only two answer choices. With stemming, different forms such as *making*, *makes*, *doing* and *does* are also provided as alternatives.

Students can use stemming to explore the variants of a word. When teaching a new word, teachers often encourage students to check its variants in a dictionary. This activity enables students to find variants and practice them by creating an appropriate quiz. For example, students use a quiz to learn more about the variants of *effect*, namely *effects*, *effective*, and *effectively*.

5.5 The *Scrambled Sentence* Exercise

The words of sentences are permuted and students sort them into their original order, to help study sentence structure. Students can select suitable material to practice on.

LLDL retrieves sentences from the digital library, according to selected criteria specified by the student:

- Word or phrases that must appear
- Corpus that the sentences come from
- Difficulty level
- Sentence type (simple, complex, or both)
- Number of sentences retrieved
- Whether to sort in ascending or descending length order.

Once the sentences have been retrieved, they are permuted and presented one after another. The search terms are put in their correct position, highlighted in blue. Students can view the title of the document containing the sentence, and the sentences preceding and following it, by clicking the *help* icon.

In this activity, students can see what other students are doing, in order to encourage them to help each other and learn from their peers' mistakes. Their names are shown (the list is updated when students log in and out); clicking a name allows you to observe how that student unscrambles a sentence by observing his word moves. Students can use chat to discuss the exercise or help each other. Teachers can also log in and observe what the students are doing, and identify and analyze their errors.

6 Evaluation

LLDL demonstrates the roles that digital libraries can play in language study. It has been extensively evaluated, although we have not attempted to assess effectiveness—whether it results in efficient learning—because this paper addresses digital library issues rather than educational ones. We have also drawn a line between evaluating the system itself and evaluating the language material that teachers have put into it.

We conducted four kinds of evaluation: metadata extraction, usability, and activity evaluation with both teachers and learners. We recruited three different kinds of evaluator: usability experts, teachers, and students. The teachers also contributed to the system throughout its development, and helped recruit language students as evaluators. The evaluation is anecdotal rather than quantitative.

6.1 Evaluating Metadata Extraction

Extracted metadata provides the underlying framework for LLDL by facilitating automatic exercise generation for the various language activities. However, they are not always accurate. Sample documents were used to assess the accuracy of sentence boundary detection and identify language constructions and collocations. We identified several tags that had been incorrectly assigned by OpenNLP, causing errors in both the *Tagged sentence* metadata and the values associated with the syntactic metadata types. Four factors affect the accuracy of *collocation* metadata. First, errors in tagging produce incorrect matches against the underlying syntactic pattern. Second, the numbers used to calculate the *t*-values are not exact. Third, the choice of the rejection threshold is arbitrary. Fourth, groups of words that commonly come together more often than chance are not necessarily good collocations.

6.2 Evaluating Usability

Evaluators examined the interface and judged its compliance with recognized usability principles. They focused on:

- Explicitness: users understand how to use the system
- Compatibility: operations meet expectations formed from previous experience
- Consistency: similar tasks are performed in similar ways
- Learnability: users can learn about the system's capability
- Feedback: actions are acknowledged and responses are meaningful.

Three rounds of usability evaluation were conducted, by usability experts, students, and language teachers. This feedback was used to improve the interface before embarking on the next stage of evaluation.

6.3 Evaluating Activities by Language Teachers

We showed the system to teachers at an early stage, and they proposed several activities that were incorporated into the system we have described. We also made other modifications based on their feedback, giving more search options for the scrambled sentence exercises, excising only nouns and verbs in the *Predicting words* activity, and showing students extracted collocations for the *Collocation identification* activity.

Later we performed a further evaluation, focusing on:

- Do the activities meet the teachers' original expectation?
- What do they think of the feedback provided to students?
- Which ability levels are the activities suitable for?
- What do they think of the exercise material that is used?

On the whole, the teachers thought the activities exceeded their original expectations. They especially liked the use of authentic reading material. They also liked the feedback provided to students, particularly the summaries provided at the end of exercises. They made many constructive and detailed comments on the individual exercises which were used for further improvements such as providing help and hints.

6.4 Evaluating Activities by Language Learners

Ten language learners, from 18 to 67 years old and native speakers of Arabic, Chinese, Italian and Japanese, participated in an experiment aimed at assessing student satisfaction with the activities. They were grouped into beginner (2), intermediate (4) and advanced (4), and paired up with like partners. In each session they tried out three activities. They filled out a questionnaire and answered verbal questions. The eight activities were allocated to the different levels in accordance with the teachers' advice.

On the whole, we did not learn a great deal from the language learners themselves. It was gratifying to find that the participants liked the activities, and appreciated any opportunity to do exercises outside the classroom. They could understand the feedback provided, but would have liked explanations of answers to be provided. It is easier for younger people with better computer skills to adjust to this learning environment and make the most of it. The evaluation also showed that the competitive activities are more attractive to younger students, and (predictably) male ones.

7 Conclusion

Digital libraries have stunning potential for improving language teaching and learning. But while there are thousands of language learning systems on the web, the potential of digital libraries in this domain remains virtually untapped. Digital libraries contain authentic text, have comprehensive search capabilities, and can automatically generate precisely-targeted exercise material. They can also provide a social environment for students to work in. Teachers can build their own collections—such as the image collections used for the image guessing activity. The library paradigm of assigning metadata to documents serves to separate the structure of exercises from their content. The digital library paradigm of automatic metadata extraction frees teachers from the onerous task of producing exercise material by hand.

We have demonstrated that stimulating educational activities can be build on top of digital libraries that have been augmented with metadata designed specifically to support language teaching. The activities are novel, and incorporate elements of cooperation, competition, and communication. All use authentic material from the digital library instead of artificial made-up examples. They have analogies in traditional classroom activities used for language teaching, but most go much farther than it is feasible to do in the classroom environment, particularly under the inevitable constraints of material that has to be carefully prepared in advance.

The activities we have devised can be used in a classroom setting or for private study. In exercises that involve pairs of students, the system matches them up automatically. In many cases students can create their own exercises. The chat facility provides a social environment that is integrated into the educational setting.

The interface to the LLDL system is explicitly designed to be multilingual (it is available in Chinese as well as English). Resource bundles for different languages have the same set of keys, which the program uses internally by the program, and different value strings for different languages. They are named following conventions that make them easy to locate. To add a new interface language, you simply create a bundle for that language and drop it into the folder where the resources are stored.

The system is not restricted to teaching English, the current target language. To extend it to other languages, difficulty metrics and open source implementations of rudimentary parsing are needed in the target language. Second language learning is one of society's greatest challenges; one that is particularly relevant to Europe. We believe language learning will prove a key application of the European digital library.

References

1. Marchionini, G. and Maurer, H. (1995) "The role of digital libraries in teaching and learning." *Comm ACM*, 38(4), 67-75.
2. Witten, I.H. and Bainbridge, D. (2003) *How to build a digital library*. Morgan Kaufmann.
3. Edge, J. (1993) *Essentials of English language teaching*. Addison Wesley Longman.
4. Clouston, M.L. (1997). Towards an understanding of culture in L2/FL education. *Ronko: K.G. Studies in English*, 25, 131-150.
5. Warschauer, M. and Healey, D. (1998) "Computers and language learning: An overview." *Language Teaching* 31: 57-71.
6. Vygotsky, L.S. (1978). *Mind and society*. Cambridge, MA: Harvard University Press.
7. Friesen, N. Mason, J. and Ward, N. (2003) "Building educational metadata application profiles." *Proc Int Conf on Dublin Core and Metadata for e-Communities*, pp. 63-69.
8. Flesch, R. (1948). A new readability yardstick. *J. Applied Psychology*, *32*, 211–233.
9. Lewis, M. (1997) *Implementing the Lexical Approach*. Language Teaching Publications.
10. Manning, C. and Schütze, H. (1999) *Foundations of Statistical NLP*. MIT Press.
11. Hill, J. (1999) "Collocational competence." *English Teaching Professional*, V. 11, pp. 3-6.

Beyond Digital Incunabula: Modeling the Next Generation of Digital Libraries*

Gregory Crane, David Bamman, Lisa Cerrato, Alison Jones, David Mimno, Adrian Packel, David Sculley, and Gabriel Weaver

Perseus Project, Tufts University

Abstract. This paper describes several incunabular assumptions that impose upon early digital libraries the limitations drawn from print, and argues for a design strategy aimed at providing customization and personalization services that go beyond the limiting models of print distribution, based on services and experiments developed for the Greco-Roman collections in the Perseus Digital Library. Three features fundamentally characterize a successful digital library design: finer granularity of collection objects, automated processes, and decentralized community contributions.

1 Introduction

Potentially massive digital libraries such as the Google Library Project [17], the Open Content Alliance [20] and the EU i2010 Initiative [21] emphasize high volume digitization based primarily on automatically generated output of page images. While digitizing page images should be the first step in any digitization project and must comprise the core of any million book library, we need as well a sound infrastructure to augment our ability to search, browse and analyze the collection. Too great an emphasis on quantity can reinforce usage models that perpetuate limits from print distribution. This paper argues for a more aggressive, but still conservative, design strategy aimed at providing customization and personalization services that go beyond limiting models based on print distribution. While customization involves the user making explicit choices about the interface or system they are using, personalization involves the system adapting itself automatically to the user [35]. We base our argument on existing services and experiments developed for the Greco-Roman collections in the Perseus Digital Library [8, 6, 7]. The underlying methods are broader in application, have contributed to work completed for the National Science Digital Library [19], and lay the foundation for a range of services in Fedora and other digital repositories [27].

This paper has the following components. First, it describes several incunabular assumptions that impose upon early digital libraries the limitations drawn

[*] The U.S. National Endowment for the Humanities (NEH) and the National Science Foundation provided primary support for this work under the Digital Library Initiative, Phase 2. Support from the Institute for Museum and Library Services also contributed substantially to this project.

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 353–366, 2006.

from print [34]. Second, it describes three features that fundamentally characterize emergent digital libraries. Third, it provides examples of customization and personalization built upon these three features.

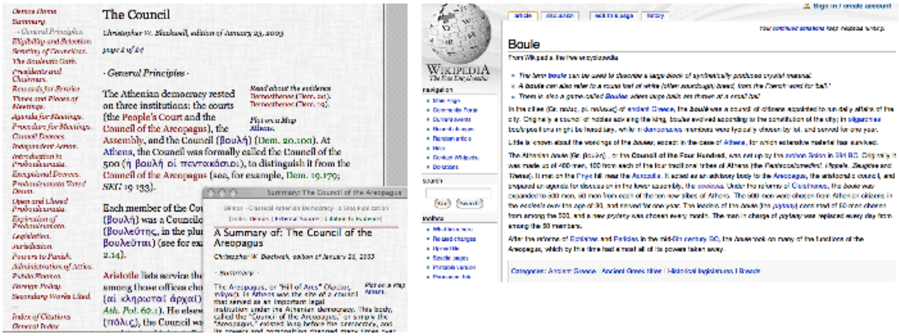
2 Incunabular Models of Early Digital Libraries

New media begin by solving well known problems but also by imitating the forms that precede them. Consider three habits of thought drawn from paper libraries that constrain the design of digital libraries. First, academic publications are based on coarse chunks of information, usually PDF or HTML files with heavily structured data (e.g., section headers, notes, bibliography). Pre-structured documents perpetuate the primacy of the author, leaving readers to do what they can with the structure and content that the author has chosen to include. Second, the emphasis on metadata carries forward the card catalogue of the print library. While metadata is important, metadata repositories that do not also include content are of limited use and have imposed an elegant modularity that constrains, as much as it has enhanced, intellectual life [26]. Third, print libraries are static. We may replace the books on the shelves with new editions, but those old books do not generate new versions of themselves. Print libraries cannot learn about their holdings and generate new content on their own. Information retrieval systems, which reindex libraries as they grow, constitute only a partial step in this direction, for they do not cycle over and over generating new knowledge by learning from their collections and from their users.

Hand-crafted collections, their contents marked up according to XML DTD/schemas with RDF metadata, incorporate major advantages over, but still narrowly replicate, their print antecedents. Some projects have, however, begun to move beyond these limitations. Figure 1 (left) is drawn from the most current and best documented survey of Athenian democracy now available: an electronic publication called Demos, published on-line as part of the Stoa publishing consortium [18].

First, Demos combines traditional and emerging structural designs: it can be downloaded as PDF chapters resembling conventional publications, but it is also available as logical chunks, each containing a semantically significant unit of text, providing greater precision than chapter heading and greater accuracy than page numbers in a book.

Second, Blackwell based a densely hypertextual work on an academic digital library that contained most of the primary evidence about Athenian democracy. The digital library shapes the form of Demos. Every major statement contains links to the primary evidence: where print reference works avoid visual clutter and save space, style sheets can turn these links on and off in a digital publication. More substantively, documented writing diminishes authorial claims of authority, offering readers an opportunity to compare conclusions with their evidence and enabling discussion. Unlike their print antecedents, citations in a digital library can point not only to pre-existing documents but also to services such as morphological analysis of Greek words or the plotting of geographic locations on high-resolution maps. The author combines links to the digital library and contextualizing materials within Demos (one of which is pictured in the figure). These internal links include discussions of the primary sources and the issues that they raise. Demos does not, however, directly address the model of the static library: each chunk of Demos lists its last modification dates, and each date testifies to the fact that Demos, for all its strengths, is not changing.



Demos Home
Summary
→ General Principles
Eligibility and Selection
Scrutiny of Councilors
The Bouleutic Oath
Presidents and Chairmen
Rewards for Service
Times and Places of Meetings
Agenda for Meetings
Procedure for Meetings
Council Decrees
Independent Action
Introduction to Probouleumata
Exceptional Decrees
Probouleumata Voted Down
Open and Closed Probouleumata
Expiration of Probouleumata
Legislation
Jurisdiction
Powers to Punish
Administration of Attica
Public Finance
Foreign Policy
Secondary Works Cited
—
Index of Citations
General Index

The Council

Christopher W. Blackwell, edition of January 23, 2003

page 2 of 24

· General Principles ·

The Athenian democracy rested on three institutions: the courts (the People's Court and the Council of the Areopagus), the Assembly, and the Council (βουλή) (Dem. 20.100). At Athens, the Council was formally called the Council of the 500 (ἡ βουλὴ οἱ πεντακόσιοι), to distinguish it from the Council of the Areopagus (see, for example, Dem. 19.179; SEG 19 133).

Read about the evidence
Demosthenes (Dem. 20).
Demosthenes (Dem. 19).

Plot on a Map
Athens.

Each member of the Cou
(βουλή) was a Councilo
(βουλεύτης, in the plu
βουλεῦται) (see for ex
2.14).

Aristotle lists service the
among those offices cho
(αἱ κληρωταὶ ἀρχαί)
Ath. Pol. 62.1). He elsew
(πόλις), the Council wa

Summary: The Council of the Areopagus

Demos - Classical Athenian Democracy - a Stoa Publication

[Index: Demos | External Sources | Citation to Evidence]

A Summary of: The Council of the Areopagus

Christopher W. Blackwell, edition of January 26, 2003

· Summary ·

The Areopagus, or "Hill of Ares" (Ἄρειος πάγος), in Athens was the site of a council that served as an important legal institution under the Athenian democracy. This body, called the "Council of the Areopagus," or simply the "Areopagus," existed long before the democracy, and its powers and composition changed many times over

Plot on a Map
Athens.

Your continued donations keep Wikipedia running!

Boule

From Wikipedia, the free encyclopedia

- The term boule can be used to describe a large block of synthetically produced crystal material.
- A boule can also refer to a round loaf of white (often sourdough) bread, from the French word for 'ball.'
- There is also a game called Boules where large balls are thrown at a small ball.

In the cities (Gr. πόλις, pl. πόλεις) of Ancient Greece, the boulē was a council of citizens appointed to run daily affairs of the city. Originally a council of nobles advising the king, boulai evolved according to the constitution of the city; in oligarchies boule positions might be hereditary, while in democracies members were typically chosen by lot, and served for one year.

Little is known about the workings of the boulai, except in the case of Athens, for which extensive material has survived.

The Athenian boule (Gr. βουλή), or the Council of the Four Hundred, was set up by the archon Solon in 594 BC. Originally it was made up of 400 men, 100 from each of the four traditional tribes of Athens (the Pentakosiomedimnoi, Hippeis, Zeugitae and Thetes). It met on the Pnyx hill near the Acropolis. It acted as an advisory body to the Areopagus, the aristocratic council, and prepared an agenda for discussion in the lower assembly, the ecclesia. Under the reforms of Cleisthenes, the boule was expanded to 500 men, 50 men from each of the ten new tribes of Athens. The 500 men were chosen from Athenian citizens in the ecclesia over the age of 30, and served for one year. The leaders of the boule (the prytany) consisted of 50 men chosen from among the 500, and a new prytany was chosen every month. The man in charge of prytany was replaced every day from among the 50 members.

After the reforms of Ephialtes and Pericles in the mid-5th century BC, the boule took on many of the functions of the Areopagus, which by this time had almost all of its powers taken away.

Categories: Ancient Greece | Ancient Greek titles | Historical legislatures | [illegible]

Fig. 1. (Left) A page from Demos: Classical Athenian Democracy, ed. Christopher Blackwell. (Right) Wikipedia article on the Athenian Boule.

Figure 1 (right) shows the Wikipedia article on the Athenian council (boule) as it appears on March 2, 2006 [14]. Taken together, these demonstrate the possibilities of both existing and emergent digital libraries. While the Demos discussion of the boule is rigorous, it is also dated (January 23, 2003) and grows slowly out of date with each passing day. Second the Demos article reflects the synthesis of a single author interacting with a small editorial community, and such a publication method could omit important perspectives from an authoritative discussion. The Wikipedia article, by contrast, is subject to constant modification. Wikipedia can thus capture broader perspectives and remain current if opinion shifts [5]. The Wikipedia article, however, contains no source citations: modeled on contemporary encyclopedias and reference works and their relatively superficial bibliographic apparatus, Wikipedia articles contain high statement/evidence ratios. They thus claim credibility, and the lack of systematic pointers to evidence is problematic. If every statement were linked to its source, gross misrepresentations would be much more readily identified and corrected.

3 Three Features That Characterize Post-incunabular Digital Libraries

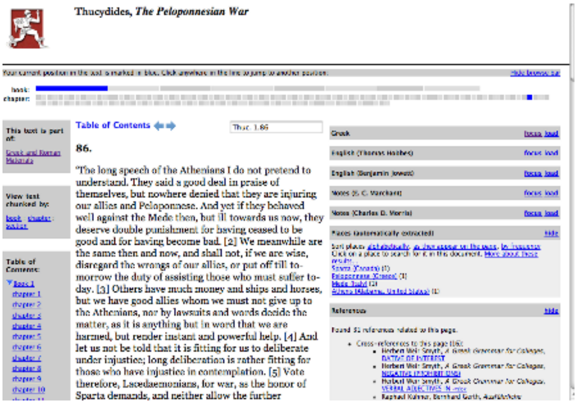
If we combine the scholarly rigor of Demos with the self-organizing qualities of Wikipedia, we can begin to see emerging a new model not only for digital libraries but also for the disciplined intellectual discourse which digital libraries should support. At least, three strategic functions distinguish emerging digital libraries from their predecessors.

1. Finer granularity: while many documents (like this paper) cite other publications as chunks, users often do not want to plough through an entire information object but would rather use an overview or particular sub-objects (proposition, bibliographic references, etc.) [37]. As digital libraries evolve, these structures go beyond those implicit in traditional publication models and begin to include explicit propositional statements.
2. Autonomous learning: Improved granularity implies this second fundamental characteristic of emergent digital libraries. Digital libraries should be constantly learning and becoming more intelligent as they scan their contents – a phenomenon already visible in rudimentary form with existing search engines. In a model digital library, the articles should be constantly scanning for new secondary sources, new editions of the primary materials and, to the extent possible, shifts in language that suggest perspectives that differ from the content of the current document. Thus documents and their subcomponents should be in constant, autonomous communication with the libraries of which they are a part, scrutinizing new materials submitted and rereading the rest of the collection as automated processes evolve [39].
3. Decentralized, real-time community contributions: Wikipedia presents one of the most important practical experiments in the history of publication. For all the criticism that it may deserve, the English language Wikipedia has generated more than 1,000,000 entries in five years and supports several million queries a day. If we go beyond the higher level prose and examine individual verifiable propositions, the accuracy is remarkable. A recent study of relational statements in Wikipedia demonstrated that 97.22% of basic propositional statements and 99.47% of disambiguating links prove to be correct [38]. Thus, even if we reject the expository prose of Wikipedia for bias, the propositions within Wikipedia demonstrate that decentralized communities will accumulate immense amounts of highly accurate propositional data [28].

The following sections describe concrete, if rudimentary, steps Perseus has taken toward all three of the above principles, and addresses the issues that digital libraries in general must confront in order to transcend the limitations of print distribution.

3.1 Granularity

Cultural heritage documents often have complex contents that do not lend themselves to simple hierarchical representation [3]. While modern publications predefine form and structure for the sake of simplifying system design, cultural heritage documents often have multiple, overlapping hierarchies (e.g., Thucydides' *History of the Peloponnesian War*, for instance, can be organized by book/chapter or by alternations of third person narrative and speeches). Digital libraries need to be able to address these various parts of a document that users want to work with. Figure 2 illustrates a dynamically generated report on what the Perseus DL knows about a particular chapter in Thucydides (Thuc. 1.86).


Thucydides, *The Peloponnesian War*
Your current position in the text is marked in blue. Click anywhere in the line to jump to another position:
hide browse bar
book:
chapter:
This text is part of:
Greek and Roman Materials
View text chunked by:
book : chapter : section
Table of Contents:
Book 1
chapter 1
chapter 2
chapter 3
chapter 4
chapter 5
chapter 6
chapter 7
chapter 8
chapter 9
chapter 10
Table of Contents
Thuc. 1.86
Greek focus load
English (Thomas Hobbes) focus load
English (Benjamin Jowett) focus load
Notes (E. C. Marchant) focus load
Notes (Charles D. Morris) focus load
Places (automatically extracted) hide
References hide

86.

"The long speech of the Athenians I do not pretend to understand. They said a good deal in praise of themselves, but nowhere denied that they are injuring our allies and Peloponnese. And yet if they behaved well against the Mede then, but ill towards us now, they deserve double punishment for having ceased to be good and for having become bad. [2] We meanwhile are the same then and now, and shall not, if we are wise, disregard the wrongs of our allies, or put off till to-morrow the duty of assisting those who must suffer to-day. [3] Others have much money and ships and horses, but we have good allies whom we must not give up to the Athenians, nor by lawsuits and words decide the matter, as it is anything but in word that we are harmed, but render instant and powerful help. [4] And let us not be told that it is fitting for us to deliberate under injustice; long deliberation is rather fitting for those who have injustice in contemplation. [5] Vote therefore, Lacedaemonians, for war, as the honor of Sparta demands, and neither allow the further

Sort places: alphabetically, as they appear on the page, by frequency
Click on a place to search for it in this document. More about these results...
Sparta (Canada) (1)
Peloponnese (Greece) (1)
Media (Italy) (1)
Athens (Alabama, United States) (1)

Found 51 references related to this page.

- Cross-references to this page (16):
  - Herbert Weir Smyth, *A Greek Grammar for Colleges*, DATIVE OF INTEREST
  - Herbert Weir Smyth, *A Greek Grammar for Colleges*, NEGATIVE PROHIBITIONS
  - Herbert Weir Smyth, *A Greek Grammar for Colleges*, VERBAL ADJECTIVES IN -τέος
  - Raphael Kühner, Bernhard Gerth, *Ausführliche*

Fig. 2. Information about Thucydides, *History of the Peloponnesian War*, book 1, chapter 86, with user focus upon one of several translations. The right hand of the screen illustrates other dynamically collected resources extracted from digital objects and organized into an ontology of document types.

First, information display depends upon an authority list of meaningful citations: "Thuc. 1.86" provides a common designation with which references to Thucydides are aligned. Using this citation, we can identify which digital objects mention this particular chunk of text. Second, the individual passages that cite this passage have also been classified. Using this classification, we can distinguish Greek source texts from English translations and annotations specifically about Thuc. 1.86, and both from texts that only mention this passage in passing. Third, all documents in the collection have been broken down into structural units. Thus, we can extract multiple Greek versions or foreign language translations of the precise chunk designated by Thuc. 1.86. This allows us to identify not only that Thuc. 1.86 shows up in a particular Greek grammar, for instance, but that it also appears specifically in the discussion on "dative of interest."

Such fine grained chunking can transform the value of information: the main Greek-English lexicon entry on the Greek preposition "pros" mentions Thuc. 1.86, but few readers would scan all fifty-two senses for the particular sense that cites Thuc. 1.86. Because we have captured the tree-structure of the dictionary, we know that this citation occurs at the third level down (as sense C.I.6), and we can extract this precise chunk from the much larger article. The right hand column aggregates and organizes these citations into a single report that could, in turn, be filtered and personalized for particular users.

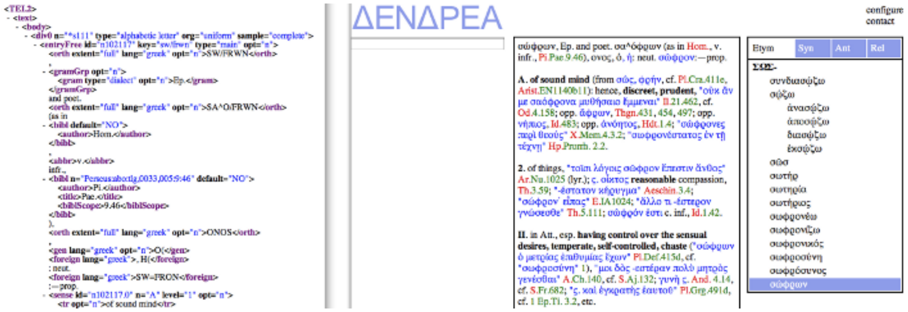
Automatically organized reports on chunks of text (or museum objects, archaeological sites or other entities) build upon well structured documents which were designed for reference and subsequent citation. In order to address the complexity of cultural heritage documents that do not lend themselves to such representation, digital libraries need to confront the following issues, none of which are glamorous but each of which demands resolution:

1. Consistent markup for complex documents: For all of the progress that has been made, we do not have large, interoperable, richly structured documents in TEI or any other markup. Capturing chapters, sections, notes and similar well-defined elements is not the problem. Rather, we need consistent ways of managing documents within documents. Some carefully marked up collections (such as DocSouth [13] and American Memory [15]) may contain indices with accurate transcriptions of citations but have not included markup that captures the structure of the index or expands often idiosyncratic abbreviations into machine actionable data.
2. Mechanisms, automated and semi-automated, to identify meaningful chunks of individual documents: This desideratum addresses the need to generate large quantities of markup scalably as digital libraries with millions of books emerge: tables, notes, address/sender/ receiver/dateline tags for embedded letters and documents, indices, etc. are difficult to extract even from well transcribed documents. Such tools need to support both markup (e.g., where they add valid elements to existing structures) and extraction (where TEI texts might provide the source for generating external ontologies). Pilot projects integrating information extraction into digital libraries have begun to emerge (GATE/Greenstone), but developed solutions will be complex for large, heterogeneous collections and much needs to be done [40].
3. Digital library systems that can represent complex documents: We need as a starting point systems that support multiple, overlapping hierarchies within the same document [9]. We also need to be able to manage partial structures (e.g., browse all quoted speech, excerpts of poetry, personal names). Digital libraries need to model complex documents early in their development rather than concentrating on structurally simpler data types.
4. Conventions for exchanging complex content: Even if we address the first three issues, we still need conventions whereby we can exchange and recombine chunks of data from multiple collections [23]. These conventions include citation schemes, standard credit lines for author, institutions and funders, and an infrastructure for redundant, robust access in the present and for preservation into the future. Figure 3 illustrates an initial version of such a service, with a Perseus dictionary entry delivered as XML fragment (left) and a third party (Dendrea) representation of that data (right), with services added that are not currently present in the home digital library of the dictionary.

3.2 Automated Processes

While granularity can give us the ability to bring together related sections from different pre-existing documents, automatic analysis gives digital libraries the opportunity to create entirely new documents rather than quietly wait for new acquisitions.

Digital libraries need to include a range of automated processes that add value to their collections, including both classification (matching pre-determined patterns) and data mining (discovering new patterns) [1, 25]. While fields such as machine learning and data mining are major topics in their own right, integrating



```
<TEI.2>
 - <text>
   - <body>
     - <div0 n="*s111" type="alphabetic letter" org="uniform" sample="complete">
       - <entryFree id="n102117" key="sw/frwn" type="main" opt="n">
           <orth extent="full" lang="greek" opt="n">SW/FRWN</orth>
           ,
         - <gramGrp opt="n">
             <gram type="dialect" opt="n">Ep.</gram>
           </gramGrp>
           and poet.
           <orth extent="full" lang="greek" opt="n">SA^O/FRWN</orth>
           (as in
         - <bibl default="NO">
             <author>Hom.</author>
           </bibl>
           ,
           <abbr>v.</abbr>
           infr.,
         - <bibl n="Perseus:abo:tlg,0033,005:9:46" default="NO">
             <author>Pi.</author>
             <title>Pae.</title>
             <biblScope>9.46</biblScope>
           </bibl>
           ),
           <orth extent="full" lang="greek" opt="n">ONOS</orth>
           ,
           <gen lang="greek" opt="n">O(</gen>
           <foreign lang="greek">, H(</foreign>
           : neut.
           <foreign lang="greek">SW=FRON</foreign>
           :—prop.
         - <sense id="n102117.0" n="A" level="1" opt="n">
             <tr opt="n">of sound mind</tr>
```

σώφρων, Ep. and poet. σα^όφρων (as in Hom., v. infr., Pi.Pae.9.46), ονος, ὁ, ἡ: neut. σῶφρον:—prop.

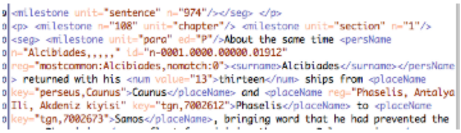
A. of sound mind (from σῶς, φρήν, cf. Pl.Cra.411e, Arist.EN1140b11): hence, **discreet, prudent,** "οὐκ ἄν με σαόφρονα μυθήσαιο ἔμμεναι" Il.21.462, cf. Od.4.158; opp. ἄφρων, Thgn.431, 454, 497; opp. νήπιος, Id.483; opp. ἀνόητος, Hdt.1.4; "σώφρονες περὶ θεούς" X.Mem.4.3.2; "σωφρονέστατος ἐν τῇ τέχνῃ" Hp.Prorrh. 2.2.

2. of things, "τοῖσι λόγοις σῶφρον ἔπεστιν ἄνθος" Ar.Nu.1025 (lyr.); ς. οἶκτος **reasonable** compassion, Th.3.59; "-έστατον κήρυγμα" Aeschin.3.4; "σώφρον' εἶπας" E.IA1024; "ἄλλο τι -έστερον γνώσεσθε" Th.5.111; σῶφρόν ἐστι c. inf., Id.1.42.

II. in Att., esp. **having control over the sensual desires, temperate, self-controlled, chaste** ("σώφρων ὁ μετρίας ἐπιθυμίας ἔχων" Pl.Def.415d, cf. "σωφροσύνη" 1), "μοι δὸς -εστέραν πολὺ μητρὸς γενέσθαι" A.Ch.140, cf. S.Aj.132; γυνὴ ς. And.4.14, cf. S.Fr.682; "ς. καὶ ἐγκρατὴς ἑαυτοῦ" Pl.Grg.491d, cf. 1 Ep.Ti. 3.2, etc.

Fig. 3. (Left) A well-formed fragment of XML representing an entry from the Liddell Scott Jones lexicon served over the Web. This service supports third party added-value services that exploit the rich underlying structure. Note that this document has a carefully captured structure, with each sense definition possessing a unique identification number. Thus, individual senses can be precisely extracted and reused in third-party hybrid documents. (Right) A third-party representation of the same dictionary article. Dendrea.org not only provides a different front end but also information extraction services that scan for etymological data, related words, and semantic relationships such as synonymy and antonymy.

such technologies into digital libraries raises a range of problems [11, 10]. We need systems that can draw upon the contents of the digital library, continually applying new knowledge sources (e.g., gazetteers, machine readable dictionaries) as these become available, recalculating its results and assessing its performance [31, 41].



```
<milestone unit="sentence" n="974"/></seg> </p>
<p> <milestone n="108" unit="chapter"/> <milestone unit="section" n="1"/>
<seg> <milestone unit="para" ed="P"/>About the same time <persName
n="Alcibiades,,,,," id="n-0001.0000.00000.01912"
reg="mostcommon:Alcibiades,nomatch:0"><surname>Alcibiades</surname></persName
> returned with his <num value="13">thirteen</num> ships from <placeName
key="perseus,Caunus">Caunus</placeName> and <placeName reg="Phaselis, Antalya
Ili, Akdeniz kiyisi" key="tgn,7002612">Phaselis</placeName> to <placeName
key="tgn,7002673">Samos</placeName>, bringing word that he had prevented the
```

Fig. 4. Named entity analysis for classical texts. Notice the lack of culturally appropriate markup, with the TEI SURNAME tag used problematically to capture the primary name, as we begin adapting instruments for modern sources to Greco-Roman documents.

Not only do we need scalable methods to identify the semantically significant document chunks such as tables, embedded letters, and notes, we need more ways to analyze raw text and classify it as propositional knowledge, bibliographic citations, quotations, and named entities (e.g., is "London" a person or a place, and, which person or place?). People, places, dates and organizations are fundamental data which any mature digital library must track. Figure 4 displays the results of a named entity recognition system in place at Perseus since 2000. In an excerpt from Thucydides 8.108 ("About the same time Alcibiades returned with his thirteen ships from Caunus and Phaselis to Samos, bringing word ..."), "Alcibiades" has been automatically annotated as a person, and "Caunus," "Phaselis" and

"Samos" have been annotated as places, two of them matched with the Getty Thesaurus of Geographic Names via TGN identification numbers.

Automated analysis also includes multi-lingual services such as cross-language information retrieval and machine translation [30, 12]. Identifying the fundamental meanings of a word is a notoriously slippery problem – human lexicographers do not agree among themselves as to what constitutes a separate word sense. One pragmatic approach involves examining translation equivalents: where translators use distinct words in the translation language, we have evidence of a substantive different meaning. Thus Table 1 displays one cluster of word meanings for the polysemous Greek word *arche*, derived from comparison of a Greek source text and five separate English translations. *Arche* can mean "empire," "government," "political office," and "beginning"; by grouping together the words that occur around it, we are able to use translations to identifyits intended sense. The six texts as a whole are aligned according to the standard citation scheme rather than at the word or sentence level (c. 42 Greek words per chunk), with sections themselves in four of the five English translations automatically aligned with the Greek original. The experiment thus explores what can be done with translations of canonical texts, with minimal extra tagging, that will populate large emerging digital libraries.

Table 1. Parallel text analysis: word clusters associated with uses of the Greek word *arche* in Thucydides (c. 150,000 words) and five English translations. Translation equivalents are underlined. The clusters capture the senses "empire," "government," "political office," and "beginning." The cluster headed "ancient" (marked in bold) captures a distinct word that happens to share the stem *arch-*.

empire	dominion	power	government
office	government	magistrates	people
dominion	power	rule	Hellenes
magistrates	Theseus	people	council
ancient	descendants	temples	Pythian
whom	beginning	pits	just
called	Zancle	Pangaeus	originally

The translation analysis points to four elements of text mining relevant to digital libraries. First, this function will improve as digital libraries grow larger, because we will have access to more and more translations of source texts into a range of languages. Second, the clustering of word groups is computationally intensive and the current algorithm is not suited to providing real time results. Third, while such exploratory data may not begin as part of the general digital library, the results of such analysis, once generated, may become a domain specific service available to those reading Greek (or similar languages for which this service is suitable). We may well find domain specific front ends, integrating data from several larger collections into a new hybrid information space designed for particular communities. Fourth, the output of the parallel text analysis is useful in its own right to human analysts, but this output also provides a foundation for cross-language information retrieval, machine translation and other multi-lingual services.

3.3 User Contributions

Every large collection contains errors and, while these are finite and may be corrected in a finite period of time, by the time original errors may be fully corrected, scholarship will have marched on, rendering bibliography, front matters, and annotations in need of revision. While some automated processes do approach perfection, even these still generate errors, and most automated processes have error rates far removed from 100%. Some processes (such as assigning a sense to a given instance of a word or analyzing the syntax of a sentence) will yield disagreement among experts.

We need to consider mechanisms to collect information from our user communities [33, 4]. In some cases, the amateurs will probably perform better than the academics: professional historians may chafe at genealogists fixated on precise identifications of people, places and organizations or antiquarians fretting about the precise buttons a particular person might have worn, but such attention to detail can be of immense value when we are refining the results of our collections.

There are two categories of contribution. First, we need annotations and reference articles that incorporate at least some full text. Wikipedia has demonstrated both immense strengths and troubling weaknesses in this area. More conservative efforts such as PlanetMath [16], based on Noosphere, have arguably produced more consistent results, but they are much more focused efforts and have created around 5,000 encyclopedia entries rather than the 1,000,000 in Wikipedia [24]. The NEH has funded "Pleaides: An Online Workspace for Ancient Geography," which will explore community created scholarly content [22]. The Perseus DL will include Wiki pages for every data object and every addressable text chunk. Our optimistic hypothesis is that community driven commentaries, created by dedicated amateurs struggling to understand the texts, may prove more useful than commentaries produced by experts: we expect many errors at first but that the churning of user responses will weed out mistakes and that Wiki commentaries will evolve into accurate instruments. We are less concerned with the potential errors than with whether such Wiki commentaries will attract a critical mass of contributors.

Second, we need to collect user feedback on propositional data: e.g., whether "London" in passage X is London, Ontario, rather than London, England; whether "saucia" is a nominative singular adjective rather than ablative; whether a certain dative is an indirect object of the verb rather than a dative of possession with a nearby noun. Propositional data does not always have a single, clearly correct answer, but we can collect alternatives and store them in a structured format.

We created two initial mechanisms to collect propositional user input, allowing users to match a particular word sense and morphological analysis appropriate to a given word in a given passage. Figure 5 illustrates the results of three processes. First, a morphological analyser generates an exhaustive list of possible analyses for the form "saucia." Second, another module examines the immediate context and the relative frequency of the forms and possible dictionary entries (if the word is lexically ambiguous), then calculates probabilities for each alternative analsysis. Accuracy of the automated disambiguation stands at 76%. Third, users can cast votes of their own, whether to reinforce the automatic analysis or to suggest an alternative.

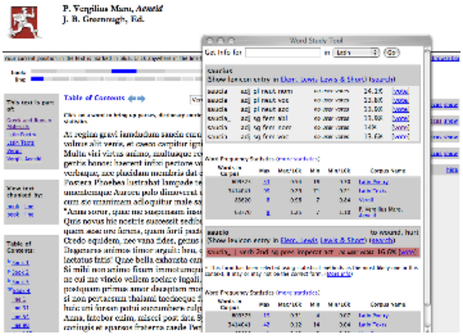
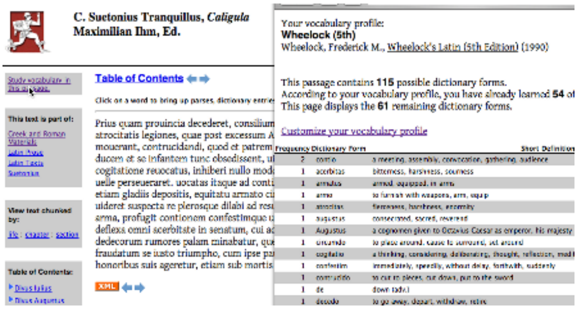

P. Vergilius Maro, Aeneid
J. B. Greenough, Ed.
Word Study Tool
Get Info for
in Latin
Go
Table of Contents
saucia
to wound, hurt
Show lexicon entry in Elem. Lewis Lewis & Short) (search)
Word Frequency Statistics (more statistics)
Corpus Name

Fig. 5. System to register votes against machine generated choice of correct morphological analysis: heuristic driven morphological analysis has been a long-term service in the Perseus DL, but we have only recently applied machine learning to estimate the correct morphological analyses for a given word in a given passage. Users can vote for alternative analyses if they disagree with the automatic choice. This system has been adopted to evaluate unfiltered, anonymous contributions, providing a possible baseline for more demanding models.

As of March 7, 2006, users have cast 7,597 votes to disambiguate Greek and Latin words with more than one possible morphological analysis. The overall accuracy for individual voting stands at 89%, but the improvement over the performance of the automated disambiguation is substantially higher, since users overwhelmingly vote on words for which the system has assigned the wrong analysis. (While the overall accuracy of automatic disambiguation is 76%, its accuracy on words that users vote on is only 34%.) And since 43% of word tokens have only one morphological sense (and do not need therefore to be disambiguated), user voting with 89% accuracy on ambiguous forms has the potential to deliver an overall system with 93.7% accuracy if every ambiguous word receives one vote. We could solicit expert users to fill in the remaining accuracy gap, but a better solution may simply be to focus on enlarging the contributor base: the more individual votes per word, the more likely that all, when taken together, will be correct.

4 User-Centered Digital Libraries: Customization and Personalization

The three incunabular constraints of early digital libraries all act as much against as for the user. The catalogue model tells users what exists and sometimes points



C. Suetonius Tranquillus, *Caligula*
Maximilian Ihm, Ed.

Click on a word to bring up parses, dictionary entrie

Prius quam prouincia decederet, consilium
atrocitatis legiones, quae post excessum A
mouerant, contrucidandi, quod et patrem
ducem et se infantem tunc obsedissent, ut
cogitatione reuocatus, inhiberi nullo mod
uelle perseueraret. uocatas itaque ad conti
etiam gladiis depositis, equitatu armato ci
uideret suspecta re plerosque dilabi ad resu
arma, profugit contionem confestimque u
deflexa omni acerbitate in senatum, cui ac
dedecorum rumores palam minabatur, que
fraudatum se iusto triumpho, cum ipse pa
honoribus suis ageretur, etiam sub mortis

Your vocabulary profile:
Wheelock (5th)
Wheelock, Frederick M., Wheelock's Latin (5th Edition) (1990)

This passage contains **115** possible dictionary forms.
According to your vocabulary profile, you have already learned **54** of
This page displays the **61** remaining dictionary forms.

Customize your vocabulary profile

Frequency	Dictionary Form	Short Definition
2	contio	a meeting, assembly, convocation, gathering, audience
1	acerbitas	bitterness, harshness, sourness
1	armatus	armed, equipped, in arms
1	armo	to furnish with weapons, arm, equip
1	atrocitas	fierceness, harshness, enormity
1	augustus	consecrated, sacred, reverend
1	Augustus	a cognomen given to Octavius Caesar as emperor, his majesty
1	circumdo	to place around, cause to surround, set around
1	cogitatio	a thinking, considering, deliberating, thought, reflection, med
1	confestim	immediately, speedily, without delay, forthwith, suddenly
1	contrucido	to cut to pieces, cut down, put to the sword
1	de	down (adv.)
1	decedo	to go away, depart, withdraw, retire

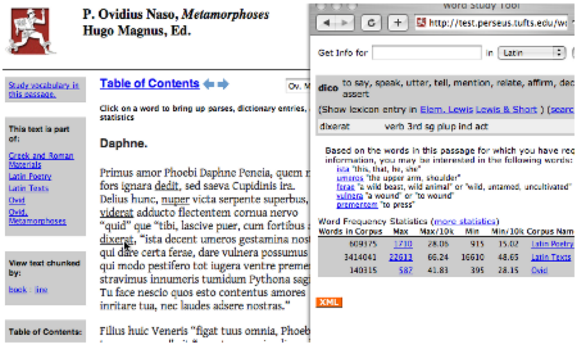
Fig. 6. Customized knowledge profile: the digital library knows the textbook with which the user has worked and analyzes probable known and unseen vocabulary in a given passage. Because the user has specified a profile and the system has responded, this is an example of customization.

them to on-line versions of the source object, but then its job is done and the users must do what they can with what they find. The digital codex model may incorporate searching and convert citations to links, but the author creates fixed content and structure around which users must work. The static library can learn neither on its own nor from its users. Early digital libraries thus, not surprisingly, replicate the hegemony of library, author, and publisher.

Digital libraries can, however, shift the balance further toward the user and toward the active life of the mind. More finely grained data objects, automated processes and decentralized user contributions all should interact, with the digital library progressively growing better structured and more self-aware. Initial data structures seed automated processes which classify and mine data. Users evaluate classification results and feed their contributions back into the system. Data mining suggests new patterns, which in turn complement or revise previous schemes, leading to the discovery and classification of new structures within the same set of digital objects.

Two fundamental strategies should be available to the user. First, users should be able to customize the environment [2]. Such customization needs to reflect not only default page layouts and simple preferences but also much more elaborate models of user knowledge. Figure 6 shows a customized report for a user who has studied Latin with a particular textbook (one of about thirty Greek and Latin textbooks whose vocabularies we have modeled on a chapter by chapter basis). Multiple readers with different backgrounds can thus see in a given passage which terms are likely to be novel and which they have encountered before. The fundamental principle can be applied to scientific terminology – which is, of course, easier to recognize than natural language.

Second, personalization augments the user-initiated decisions of customization: digital libraries should be able to analyze user behavior and background and offer new automatically generated configurations of information [36, 29, 32]. Figure 7 illustrates a recommender system that compares records of questions from earlier readers who had read a particular text. By mining past behaviors,



P. Ovidius Naso, *Metamorphoses*
Hugo Magnus, Ed.

Click on a word to bring up parses, dictionary entries, statistics

Daphne.

Primus amor Phoebi Daphne Peneia, quem n
fors ignara dedit, sed saeva Cupidinis ira.
Delius hunc, nuper victa serpente superbus,
viderat adducto flectentem cornua nervo
"quid" que "tibi, lascive puer, cum fortibus a
dixerat, "ista decent umeros gestamina nos
qui dare certa ferae, dare vulnera possumus
qui modo pestifero tot iugera ventre preme
stravimus innumeris tumidum Pythona sag
Tu face nescio quos esto contentus amores
inritare tua, nec laudes adsere nostras."

Filius huic Veneris "figat tuus omnia, Phoeb

dico to say, speak, utter, tell, mention, relate, affirm, dec assert

(Show lexicon entry in Elem. Lewis Lewis & Short) (searc

dixerat verb 3rd sg plup ind act

Based on the words in this passage for which you have req information, you may be interested in the following words:
ista "this, that, he, she"
umeros "the upper arm, shoulder"
ferae "a wild beast, wild animal" or "wild, untamed, uncultivated"
vulnera "a wound" or "to wound"
prementem "to press"

Word Frequency Statistics (more statistics)

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Nam
609375	1710	28.06	915	15.02	Latin Poetry
3414041	22613	66.24	16610	48.65	Latin Texts
140315	587	41.83	395	28.15	Ovid

Fig. 7. Personalized knowledge profile: the digital library does not know the background of the user but has analyzed four initial questions, compared these with past question patterns and suggested which of the remaining three hundred words the current user most likely to query. User behavior clusters into distinct classes and this approach has been able to predict 67% of subsequent queries. Because the system has taken the initiative rather than the user, this is an example of personalization.

the system can quickly learn to predict most of the subsequent questions that new users will pose.

5 Conclusion

Google with its massive library project, more recently Microsoft, Yahoo and others in the Open Content Alliance, and potentially the EU in its i2010 initiative are poised to assemble massive, but coarse, libraries of digitized books that have the potential to reinforce usage models based on print distribution. This paper provides initial examples of a post-incunabular design strategy utilized at the Perseus Project for its Greco-Roman collection but, we hope, scalable to other domains, focused on the principles of customization and personalization built upon fine grained digital objects, automated processes and decentralized user contributions.

References

[1] H. S. Baird, V. Govindaraju, and D. P. Lopresti. Document analysis systems for digital libraries: Challenges and opportunities. In *Document Analysis Systems*, pages 1–16, 2004.

[2] N. Beagrie. Plenty of room at the bottom? Personal digital libraries and collections. *D-Lib Magazine*, 11(6), 2005. http://dlib.anu.edu.au/dlib/june05/beagrie/06beagrie.html.

[3] J. Bradley. Documents and data: Modelling materials for humanities research in XML and relational databases. *Literary and Linguistic Computing*, 20(1), 2005.

[4] M.A.B. Burkard. Collaboration on medieval charters–Wikipedia in the humanities. In *Proceedings of the XVI International Conference of the Association for History and Computing*, pages 91–94, 2005.

[5] D. J. Cohen and R. Rosenzweig. Web of lies? Historical knowledge on the internet. *First Monday*, 10(12), Dec. 2005. http://www.firstmonday.org/issues/issue10_12/cohen/.

[6] G. Crane. Cultural heritage digital libraries: Needs and components. In *ECDL*, Rome, Italy, 16-18 Sept. 2002.

[7] G. Crane, R. F. Chavez, A. Mahoney, T. L. Milbank, J. A. Rydberg-Cox, D. A. Smith, and C. E. Wulfman. Drudgery and deep thought: Designing a digital library for the humanities. *Communications of the ACM*, 44(5):35–40, 2001.

[8] G. Crane, C. E. Wulfman, L. M. Cerrato, A. Mahoney, T. L. Milbank, D. Mimno, J. A. Rydberg-Cox, D. A. Smith, and C. York. Towards a cultural heritage digital library. In *JCDL*, pages 75–86, Houston, TX, June 2003.

[9] A. Dekhtyar, I. E. Iacob, J. W. Jaromczyk, K. Kiernan, N. Moore, and D. C. Porter. Support for XML markup of image-based electronic editions. *International Journal on Digital Libraries*, 6(1):55–69, Feb. 2006.

[10] J. S. Downie, J. Unsworth, B. Yu, D. Tcheng, G. Rockwell, and S. J. Ramsay. A revolutionary approach to humanities computing?: Tools development and the D2K data-mining framework. In *Annual Joint Conference of The Association for Computers and the Humanities & The Association for Literary and Linguistic Computing*, 2005.

[11] F. Esposito, D. Malerba, G. Semeraro, S. Ferilli, O. Altamura, T. M. A. Basile, M. Berardi, M. Ceci, and N. Di Mauro. Machine learning methods for automatically processing historical documents: From paper acquisition to XML transformation. In *DIAL*, volume 1, pages 328–335. IEEE Computer Society, 2004.

[12] F. C. Gey, N. Kando, and C. Peters. Cross-language information retrieval: the way ahead. *Information Processing and Management*, 41(3):415–431, 2005.

[13] http://docsouth.unc.edu/.

[14] http://en.wikipedia.org/wiki/Boule.

[15] http://memory.loc.gov/ammem/index.html.

[16] http://planetmath.org/.

[17] http://print.google.com/googleprint/library.html.

[18] http://seneca.stoa.org/projects/demos/article_council?page=1&greekEncoding=UnicodeC.

[19] http://www.nsdl.org.

[20] http://www.opencontentalliance.org/.

[21] http://www.theeuropeanlibrary.org/portal/index.htm.

[22] http://www.unc.edu/awmc/pleiades.html.

[23] Y. E. Ioannidis, D. Maier, S. Abiteboul, P. Buneman, S. B. Davidson, E. A. Fox, A. Y. Halevy, C. A. Knoblock, F. Rabitti, H. J. Schek, and G. Weikum. Digital library information-technology infrastructures. *International Journal on Digital Libraries*, 5(4):266–274, 2005.

[24] A. Krowne. Building a digital library the commons-based peer production way. *D-Lib Magazine*, 9(10), 2003. http://www.dlib.org/dlib/october03/krowne/10krowne.html.

[25] A. Krowne and M. Halbert. An initial evaluation of automated organization for digital library browsing. In *JCDL*, pages 246–255, New York, NY, USA, 2005. ACM Press.

[26] C. Lagoze, D. B. Krafft, S. Payette, and S. Jesuroga. What is a digital library anymore, anyway? Beyond search and access in the NSDL. *D-Lib*, 11(11), 2005. http://www.dlib.org/dlib/november05/lagoze/11lagoze.html.

[27] Carl Lagoze, Sandy Payette, Edwin Shin, and Chris Wilper. Fedora: an architecture for complex objects and their relationships. *Int. J. Digit. Libr.*, 6(2):124–138, 2006.

[28] M. Lesk. The qualitative advantages of quantities of information: Bigger is better. *J. Zhejiang Univ. Science*, 11(6A), 2005.

[29] E. J. Neuhold, C. Niederée, and A. Stewart. Personalization in digital libraries: An extended view. In *ICADL*, 2003.

[30] D. W. Oard. Language technologies for scalable digital libraries. In *International Conference on Digital Libraries*, 2004. Invited Paper.

[31] G. Pant, K. Tsioutsiouliklis, J. Johnson, and C. L. Giles. Panorama: Extending digital libraries with topical crawlers. In *JCDL*, pages 142–150, New York, NY, USA, 2004. ACM Press.

[32] M. E. Renda and U. Straccia. A personalized collaborative digital library environment: a model and an application. *Information Processing and Management*, 41(1):5–21, 2005.

[33] M. Richardson and P. Domingos. Building large knowledge bases by mass collaboration. In *K-CAP*, pages 129–137, New York, NY, USA, 2003. ACM Press.

[34] M. Riva and V. Zafrin. Extending the text: digital editions and the hypertextual paradigm. In *HYPERTEXT*, pages 205–207, New York, NY, USA, 2005. ACM Press.

[35] John Russell. Making it personal: information that adapts to the reader. In *SIGDOC '03: Proceedings of the 21st annual international conference on Documentation*, pages 160–166, New York, NY, USA, 2003. ACM Press.

[36] A. F. Smeaton and J. Callan. Personalisation and recommender systems in digital libraries. *International Journal on Digital Libraries*, 5:299–308, 2005.

[37] E. S Villamil, C. González Muñoz, and R. C. Carrasco. XMLibrary search: an XML search engine oriented to digital libraries. *Lecture Notes in Computer Science - Research and Advanced Technology for Digital Libraries*, 3652:81–91, 2005.

[38] Gabriel Weaver, Barbara Strickland, Alison Jones, and Gregory Crane. Quantifying the accuracy of relational statements in Wikipedia: A methodology. In *To appear in JCDL 06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, 2006.

[39] R. Witte. An integration architecture for user-centric document creation, retrieval, and analysis. In *Proceedings of the VLDB Workshop on Information Integration on the Web*, 2004.

[40] I. H. Witten, K. J. Don, M. Dewsnip, and V. Tablan. Text mining in a digital library. *International Journal on Digital Libraries*, 4(1):56–59, 2004.

[41] Z. Zhuang, R. Wagle, and C. L. Giles. What's there and what's not?: Focused crawling for missing documents in digital libraries. In *JCDL*, pages 301–310, 2005.

Managing and Querying Video by Semantics in Digital Library*

Yu Wang, Chunxiao Xing, and Lizhu Zhou

Department of Computer Science and Technology, Tsinghua University
wangyu02@mails.tsinghua.edu.cn, {xingcx, dcszlz}@tsinghua.edu.cn

Abstract. Management of video data is an indispensable part of digital library. However, currently most digital library systems only provide the functionality of retrieving video data by meta-data which can not fulfill users' requirements. This is due to the lack of appropriate video semantic model and powerful query interface. In this paper, we propose such a model named SemTTE together with its query language VSQL. The model incorporates features of temporal structure and typed events of video contents and organizes the whole video into a tree of events. It is implemented based on XML technology with schema and instance mapped to DTD and XML documents, and queries transformed to XQuery for evaluation.

1 Introduction

Digital Libraries have emerged as large scale and distributed information and knowledge environment and infrastructure to bring together collections, services, and people in support of the full life cycle of creation, dissemination, use, storage, and preservation [14]. Managing video data and its content is an indispensable part of digital library and is becoming more and more important. However, in most current digital library systems, video data can only be queried by meta-data, which is too limited to satisfy users. For example, queries like "get all NBA shots in which YaoMing passes the ball to McGrady" can not be answered. This kind of information is called semantic information which describes "the story of the video" from the viewpoint of real world by using objects (including conceivable objects and abstract objects), events, and relations between them.

In order to query videos by semantics, an appropriate semantic model is needed to support semantic information based annotation and retrieval which is the purpose of this paper. Once we have such a model, semantics can be annotated, stored, and later queried by users. However, existing video semantic models view video data as a sequence of mutually independent segments with related semantics, without exploring relationships between semantics of different segments. However, this kind of information is of great concern for information validation and annotation assistance. The semantic model in this paper, SemTTE, overcomes these shortcomings by considering following properties:

* Supported by the National Natural Science Foundation of China under Grant No. 60473078.

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 367 – 378, 2006.

- *Typed events*: events in video can be abstracted to a finite set of types and the video can be viewed as a sequence of instances of these types. For example, “Foul” can be an event type in NBA game and “O’Neal fouled YaoMing at …” is an instance of type “Foul”.
- *Temporal structures*: the occurrences of events are temporally ordered from the beginning to the end of a video. Each event is related to a video segment to indicate its occurrence. Constraints can be defined on the ordering and relationship of events of various types, for example, in a NBA game, there must be a first section followed by a second section.

Querying ability is also a very important aspect of a semantic model since the ultimate purpose of designing such a semantic model is to help users to find what they need efficiently and accurately.

This paper is structured as follows. Section 2 is a review of related work. Section 3 presents SemTTE with its framework and query language VSQL. For more about SemTTE’s constraint mechanism, please refer to [12].Section 4 give the implementation of SemTTE based on XML technologies. Section 5 is the conclusion.

2 Related Work

MPEG7[13] is a famous standard which provides descriptors to describe the content of video data. However, it is not a data model in the sense of database domain, thus different from what we consider in this paper. First, its purpose is to describe and exchange video content through the web; second, the semantic related functionality provided is very limited; and third, it does not provide mechanisms to support semantic-related queries.

Many models [1-10] have been proposed for managing semantics in video data. [11] provides a set of evaluation criteria for them and evaluated them according to these criteria.

Typed events and temporal relationship have been considered in some models. However, event types are only used to distinguish events with different properties, not related to temporal structure. Temporal relationships are only optional features and no mechanism is provided to define constraints about it.

For query interface, most models provide declarative languages [2-8,10], some use graphical interface[9] (query language is absent in [2,4], and in [6,7] only query processing algorithms are given). Although graphical interface is more straightforward, some queries are difficult to express, such as statistical query, and complicated condition. Some languages are too complicated for users to use. For example, in [3,5] queries are defined in a logical implication like form.

As far as query ability is concerned, most models support temporal queries and query by attributes or roles and the result may be semantic information or video sequence (in [6,8] only video segments can be returned). However, some important features are poorly supported. For example, only [10] support aggregation query; only [5,6,9] support browsing, with the limitation of only browse by scenes or clusters, not by semantics.

In summary, there are much room for enhancement for both representation and query ability. Compared with existing video semantic models, the distinguishing properties of SemTTE are as following:

- ♦ Temporal relationships between events are explicitly represented and mandatory. All event instances in a video are organized into a tree structure thus the semantic of video can be clearly represented;
- ♦ A powerful query language is provided which supports not only temporal queries and query by attributes or roles, but also aggregation queries, sub queries and joint queries among multiple domains;
- ♦ Implementation based on XML technologies is provided.

3 Video Semantic Model: SemTTE

In this section, a semantic model SemTTE will be presented which considers the previously mentioned properties: typed events and temporal structure. Due to space limitation, this paper will focus on SemTTE's schema, instance and query language. For more about constraint mechanism, please refer to [12].

In the rest of this paper, we will use NBA game domain as the example to illustrate how semantic information are represented and queried. In this domain, users are interested in events such as scoring, foul, etc. These events are temporally organized into sections and games and may be performed by players or teams. Queries may be issued by designating information about events or their performers and the result may be video segments or semantic information.

3.1 Basic Concepts of SemTTE

In SemTTE, video semantics are represented by following concepts:

Events
Events have three specific properties. First, every event occurs at some time. For example, in NBA game, a scoring event may occur at the 10th minutes and last for 5 seconds. Once a uniform time base is given, the temporal relationship of any two events can be decided. Second, an event may be composed of other events and the duration of the composite event will covers that of its component events. For example, an event "Party" may be composed of events such as "Dancing", "Drinking", or "Talking", and the duration of "Party" should cover that of the other events. Third, each event may have participants and each participant may take one or more roles. For example, a foul in a NBA may have two participants who take the roles of "fouler" and "fouled" respectively.

Entities
In SemTTE, entities are used to describe events' participants. In real world, all kinds of things, objects, or abstract concepts may participant some events and take some roles. Thus all these things may be modeled as entities in SemTTE.

Attributes
Events and entities can be meaningfully described by attributes. For example, a player may be described by the attributes name, age, and height; a team may be described by the attributes name, year of foundation, and place; an event scoring may be described

by the attribute score. The value of an attribute may be a single value, a set of values, or a list of values. For entity, the attribute or attribute set that may uniquely identify the entity is called "Entity Key". For event, the difference between roles and attributes are that attributes describe properties of events while roles describe interactions between events and entities.

Entity and Event Types
In SemTTE, types and instances are distinguished. Entities and events sharing common properties are generalized by entity types and event types with a set of attributes (or roles) representing the common properties. For example, the entity type in NBA domain may be "Player" or "Team" and the event type may be "Game", "Scoring", or "Foul". A player "YaoMing" is an instance of "Player", while "Kaman fouls YaoMing at the forth section" is an instance of "Foul".

3.2 Schema

The schema of SemTTE is composed of a set of types for entities and events.

The name of an entity together with its attributes define an *entity type*. The grammar is as following:

CREATE ENTITY *EntityTypeName* (*Attribute*, *Attribute*, ...)

While each attribute definition includes the name and domain of the attribute. Entity types for NBA game domain may be

CREATE ENTITY Player (<u>name: string</u>, age: integer, team: Team)
CREATE ENTITY Team (<u>name: string</u>, year: integer, players: {Player})

The attributes underlined are the entity keys. The attribute "players" of type "Team" is a set of Players.

The definition of an event type includes type name, role set, optional attribute set, and optional direct component type set, with following grammar:

CREATE EVENT *EventTypeName*(*RoleSet?*; *AttributeSet?*; *CompSet?*)

The element *RoleSet* and *AttributeSet* are the set of role definitions and attribute definitions respectively. The element "CompSet" is a set of event type names designating the *direct component event types*. Instances of these types are called *direct component events*. For an event instance *ei* of type *e*, only instances whose types are in "CompSet" of *e* may be component events of *ei*. Event types that are not component of any other types are called *root event types*, and their instances are called *root events*. There may be multiple root event types in a schema, with each one representing a kind of video structure.

Event type definitions in NBA games may be as following:

CREATE EVENT Game (host: Team, guest: Team;
time: Datetime; Section)
CREATE EVENT Section(;secNum: integer, hscore: integer, gscore: integer;
Scoring, Foul, FreeThrow)
CREATE EVENT Scoring (scorer: Player; score: integer)
CREATE EVENT Foul (fouler: Player, fouled: Player)
CREATE EVENT FreeThrow (shooter: Player)

With these definitions we can see that the root event type is Game which has two roles host and guest, and one attribute time. Its direct component type may be Section.

The event type Section has no roles, three attributes: number, hscore, and gscore, and its direct component type may be Scoring, Foul, and FreeThrow. The attribute "secNum" stores the number of the section, while the two attributes "hscore" and "gscore" store the score of the host and the guest respectively. Event type Scoring, Foul, and FreeThrow have no component event types, so instances of these types may not have component events.

3.3 Instance

The database instance is composed of entity instances and event trees.

All entity instances are organized into a set while each instance may be viewed as a set of attribute values. Figure 1 shows an example of entity instances in NBA domain.

Players

Name	age	team
YaoMing	26	Rockets
McGrady	27	Rockets
Kaman	24	Clippers

Teams

name	year	players
Rockets	1967	{YaoMing, McGrady}
Clippers	1970	{Kaman}

Fig. 1. Entity instances in NBA domain

For each video file (or multiple files having an integral meaning, such as a game in NBA domain or a movie in Movie domain), all event instances occurring in it are organized as a tree (called event tree). All event instances in the database forms a set of event trees. The meaning of an event tree is explained as following:

- Each node in the tree corresponds to an event instance and the root node corresponds to a root event. For each event tree there exist a one-to-one mapping from tree nodes to event instances (the function *event(n)* and *node(e)* are used to get the corresponding event instance and tree node respectively);
- The parent-child relationship between nodes corresponds to the direct composing relationship between events. For any node *n* in the tree, following constraints must be satisfied:

$$\forall nc \in getChildNodes(n) : (event(nc) \in getDirectComponentEvents(event(n)))$$

where *getChildNodes* means get the set of child nodes of node *n*, and *getDirectComponentEvents* means get the set of direct component events of an event instance.

- Order exists among sibling nodes indicating temporal ordering of events;
- Each event instance is related to a video segment. The segment related to the parent node must cover all segments related to its child nodes.

Figure 2 is an example of part of an event tree in NBA domain. Each node represents an event instance and the text beside the node is the description of the instance, including event type, attributes and roles. The root node is a "Game" instance. It has four "Section" instances as direct component events. The instance "Third section" has

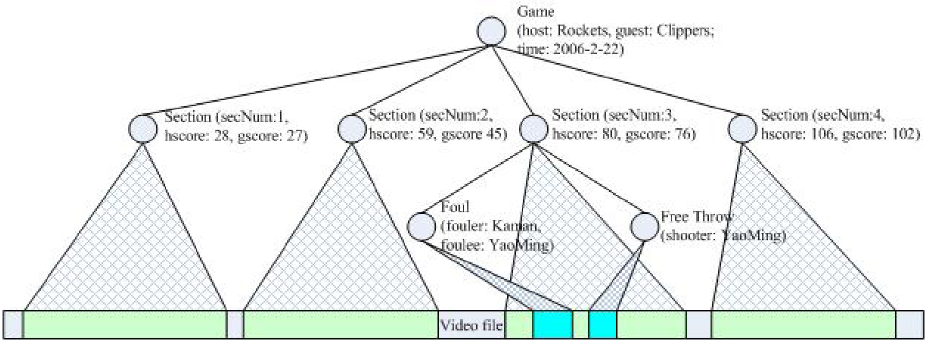


Fig. 2. Example event tree in NBA domain

two direct component events: Kaman's foul and YaoMing's free throw. The rectangle at the bottom of the figure represents the video file and each event instance in the tree is related to a video segment in this file (the video segment related to the root node is not shown for clearness) where these event instances occur.

3.4 Query

3.4.1 Query Language VSQL

In this section a powerful query language VSQL (Video Semantics Query Language) is provided which employs similar syntactical form as SQL:

```
SELECT      [WITH SEGMENT] <target list>
FROM        <variable list>
[WHERE       <search condition>]
 [GROUP BY    <target list> [HAVING <search condition>]]
[ORDER BY     <target list>]
```

The result of such a query may be video segments or semantic information consisting of targets defined in the **SELECT** clause. If more than one target follows the **SELECT** keyword, they are separated by commas. The **FROM** clause lists the variables over which the query is performed. The **WHERE** clause states a predicate that has to be satisfied by the targets in order to be included in the result set. The **GROUP BY** clause facilitates the partition of the result into subgroups so aggregate operators can be employed on each partition. The **ORDER BY** clause allows users to retrieve targets in a special user-defined order based on some target values.

Although VSQL has a similar syntax with SQL, the meanings of them have great difference. First, VSQL uses variable in the **FROM** clause instead of tables; second, VSQL uses attribute expressions in the **SELECT**, **GROUP**, and **ORDER** clause instead of column names; third, the operands in the search condition may be attribute expressions, role expressions, and variables instead of column names. Besides traditional comparison operators and set operators, new operators such as sequence operators, temporal operators etc. are introduced to deal with specific properties of video data; finally, in VSQL attributes, roles, and referenced entities or events are all accessed in object-oriented manner via the operator ".".

Now we will first present the new operators introduced for video applications, then illustrate how VSQL may be used and what the result will look like via examples. These new operators are:

- component operators: indicating composing relationship of two events, may be *DCOF* and *COF* which means "direct component of" and "component of" respectively;
- temporal operator: indicating temporal relationship between events, may be *before*, *after*, *ibefore*, and *iafter*. *before* and *after* mean that an event instance occurs before or after another event instance, but not cover or covered by it. *ibefore* and *iafter* mean that an event instance occur immediately before or after another event instance (no intermediate event instances);
- participate: The operator *participate* is used to represent an entity's participation in events. With this operator, users can find all events or all events of a specific type that an entity participates;
- is: The operator *is* is used to check whether two references (such as role expression or variable) refer to the same instance.

If the optional keyword **WITH SEGMENT** is presented in the **SELECT** clause, then the related video segment of target events will be returned.

3.4.2 Query Examples

In this section we will show the ability of VSQL more intuitively by giving some examples. Obviously, these examples can not cover all its capability. We suppose there are NBA game videos and movies in the database. Keywords are in bold and type names are in italic.

Example 1. a simple select
Return video segments, the score, and the name of the scorer of all scorings that the score greater than 1.

SELECT WITH SEGMENT s.scorer.name, s.score
FROM *Scoring* s
WHERE s.score > 1

"*Scoring* s" is a variable definition while "*Scoring*" is an event type name and "s" is the variable name. "s.score" is an attribute expression referring to the attribute "score" of "s". "s.scorer.name" is also an attribute expression with the difference that it refers to the attribute "name" of the entity that participates in the event "s" and takes the role "scorer". The result is in the following form:

s.scorer.name	s.score	Video segments
YaoMing	2	▶
Kaman	3	▶
YaoMing	2	▶
McGrady	2	▶

The icon ▶ indicates video segment and can be played by clicking it.

Example 2. a select with **GROUP**
Now we can group the previous results by players' names to show the scoring of each player.

```
SELECT WITH SEGMENT s.scorer.name, s.score
FROM Scoring s
WHERE s.score > 1
GROUP BY s.scorer.name
ORDER BY s.scorer.name, s.score
```

Now the result will look like:

s.scorer.name	s.score	Video segments
Kaman	3	▶
McGrady	2	▶
YaoMing	2	▶
	3	▶

or we can get the total score that each player gets

```
SELECT WITH SEGMENT s.scorer.name, sum(s.score)
FROM Scoring s
WHERE s.score > 1
GROUP BY s.scorer.name
ORDER BY s.scorer.name
```

and the result will be

s.scorer.name	sum(s.score)	Video segments
Kaman	3	▶
McGrady	2	▶
YaoMing	4	▶ ▶

We can see that there are two video segments in the record for YaoMing. If **WITH SEGMENT** is presented and a record involves multiple event instances, then the content of the column "Video segments" will be a list of video segments.

Example 3. a select with component operator
Get all attributes and the video segment of YaoMing's scoring in the third section of a game between Rockets and Clippers at 2006-2-22

```
SELECT WITH SEGMENT s.*
FROM Game g, Section se, Scoring s
WHERE g.time = "2006-2-22" AND s.scorer.name="YaoMing" AND
          ((g.host.name = "Rockets" AND g.guest.name = "Clippers")
            OR (g.host.name = "Clippers" AND g.guest.name = "Rockets"))
          AND s COF se AND se COF g AND se.secNum=3
```

The special symbol * denotes that the result should contain all attributes of a variable. The component operator **COF** means that s is a component event of g (s occurs during g).

Example 4. a select with "participate" operator
List all foul events that YaoMing participates in

```
SELECT WITH SEGMENT f.*
FROM Player p, Foul f
WHERE p.name = "YaoMing" AND p participate f
```

Example 5. a select with temporal operator and "is" operator
In the game at 2006-2-22, there are some players less than 25 years old who have fouled and followed immediately by a free throw. List all the fouler's name and the shooter's name, with the related video segments.

```
SELECT WITH SEGMENT p.name, fs.shooter.name
FROM Game g, FreeThrow fs, Foul f, Player p
WHERE g.time = "2006-2-22"
          AND fs COF g        AND f COF g
          AND f.fouler is p   AND p.age < 25
          AND f ibefore fs
```

Example 6. a select with sub query and involving multiple domains
Show the segments of all events in the movie domain involving a NBA star who had got more than 30 scores in a game at 30 December, 2005.

```
    SELECT WITH SEGMENT e.*
    FROM Movie m, Event e, Player p
    WHERE p participate e AND e COF m
         AND p in ( SELECT pl
              FROM Player pl, Game ga, Score s
              WHERE ga.time = “2005-12-30” AND s.scorer is pl
                   AND s COF ga
              GROUP BY pl HAVING sum(s.score) > 30)
```

4 Implementation of SemTTE

Since SemTTE is a logical model, its implementation can be based on existing technologies. After comparison among relational database, object-oriented database and XML database, we finally choose XML database. The reasons lie in following aspects. Firstly, XML is a semi-structured data model and a XML document can be viewed as a tree with ordering among elements, so the structure of event tree can be mapped to XML documents straightforwardly. Secondly, a powerful and flexible query language XQuery is provided for XML. By using the FLWOR statement, complex queries can be issued and the result may be arbitrarily structured. Finally, XML is a widely used standard to store and exchange information. Thus by choosing XML database as the underground implementation, semantics of videos can be widely shared and exchanged.

4.1 Mapping of Schema and Instance to XML DTD and Documents

The schema of SemTTE corresponds to a set of DTDs in the XML database. All entity types are stored in one DTD and all event types are first clustered (each cluster contains one and only one root event type, and all other types are direct or indirect component types of the root type) then each cluster is mapped to a DTD. The database instance corresponds to a set of XML documents. All entity instances are stored in one XML document and each event tree is stored in one XML document. Following are parts of the two XML documents generated from the entity instances and event tree in section 3.3.
Part of the entity document:

```
<EntitySet>
  <Team ID="n1" name=”Rockets” year=”1967” players=”YaoMing;McGrady”
                  playersID=”n3;n4”/>
  <Team ID="n2" name=”Clippers” year=”1948” players=”Kaman”
                  playersID=”n5”/>
  <Player ID="n3" name=”YaoMing” age=”26” team=”Rockets” teamID=”n1”/>
  <Player ID="n4" name=”McGrady” age=”27” team=”Rockets” teamID=”n1”/>
  <Player ID=”n5” name=” Kaman” age=”27” team=”Clippers” teamID=”n2”/>
</EntitySet>
```

Part of the event tree document:

```
<EventTree ID="v9">
 <Game ID="v10" time=”2006-2-22” host=”Rockets” hostID=”n1”
                guest=”Clippers” guestID=”n2”>
   <Section ID="v4" num=”3” hscore=”80” gscore=”76”>
     <Foul ID=”v5” fouler=”Kaman” foulerID=”n5”
                fouled=”YaoMing” fouledID=”n3”/>
     <FreeThrow ID="v6" shooter=”YaoMing” shooterID=”n3”/>
   </Section>
 </Game>
</EventTree>
```

As we can see, attributes and roles are mapped to element attributes whose values are strings generated from the value of the attribute or role. Values of atomic types are transformed to strings straightforwardly (the attribute “name”, “age”, etc.). And for entity references, both the identifier and the value of entity key of the referred entity are stored (such as “fouler” and “foulerID” of Foul) With this strategy, the number of join operations needed to evaluate queries on these XML documents can be reduced dramatically, thus reducing the response time. Although this will introduce redundancy, the overall efficiency will not be harmed since in video applications more than 95% of users’ requests are querying, and updating requests are very seldom.

4.2 Query Evaluation

Evaluation of VSQL query is composed of three steps. First the pre-processing will check the query to make sure that all expressions are valid and conditions in the **WHERE** clause are not contradictive. Then the query will be mapped to XQuery with the help of schema and evaluated. If the keyword **WITH SEGMENT** is presented, event instances’ temporal locations will also be returned. Finally, the post-processing will get the corresponding video segments according to the temporal locations and generate the result table.

In the second step, multiple XQuery queries can be generated from a VSQL query. Among these queries, there usually exists one that is more efficient than the others. Although whether a query is efficient is related to the optimization strategies utilized in the implementation of XQuery, a well generated query is more likely to be evaluated efficiently and less dependant on the implementation of XQuery. For example, for VSQL query:

SELECT f.*
FROM *Player* p, *Foul* f
WHERE p.age <25 **AND** p **participate** f

two XQuery queries may be generated as following:

```
for $p in collection(“NBA”)//Player,
    $f in collection(“NBA”)//Foul
where $p/@age < 25
      AND ( f/@foulerID=$p/@ID OR f/@fouledID=$p/@ID)
return …
```

```
and:
for $p in collection("NBA")//Player[@age < 25],
  $f in
   collection("NBA")//Foul[@foulerID=$p/@ID or @fouledID=$p/@ID]
return …
```

In the second query, the sizes of the intermediate result set of evaluation of the XPath expression for variable "$p" and "$f" are much smaller than that in the first query, thus it will be evaluated faster than the first one. In our implementation, following strategies are utilized to produce more efficient XQuery queries:

- ♦ All comparison expressions that do not involve join operations explicitly or implicitly are extracted from the **WHERE** clause and combined with the variable to form the variable definition in XQuery.

  For example, in the previous query, the comparison expression "p.age < 25" is extracted from the **WHERE** clause to form the variable definition in XQuery for "p" as "for $p in collection("NBA")//Player[@age < 25]".
- ♦ The comparison expression involving component operators are extracted from the **WHERE** clause to be included in the variable definition in XQuery.

  For example, the corresponding variable definition in XQuery of variable "f" in Example 5 in section 3.4.2 is "$f in $g//Foul" instead of "$f in collection('')//Foul" plus a condition "exists($g//node() intersect $f)" in the where statement.
- ♦ To utilize the previous two strategies, the condition in **WHERE** clause should first be transformed into conjunctive normal form, and if necessary the VSQL query will be decomposed to several queries to make sure that in the condition of each query the above comparison expressions can be extracted. The common part of these queries will be stored as the intermediate result to prevent it from being evaluated for multiple times.

5 Conclusion and Future Work

In this paper, a video semantic model SemTTE is proposed to manage semantics of videos. It considers *temporal structure* and *typed events* of videos and organizes the whole video into a tree of events. A powerful query language VSQL is proposed which supports not only query by attributes or roles but also aggregation queries, sub query etc. XML is chosen as the underlying implementation, so video semantics can be easily shared and exchanged.

References

[1] Adah, S.; S.Candan, K.; Chen, S.-s. The Advanced Video Information System: Data Structures and Query Processing. ACM Multimedia Systems 1996, 4, 172-186

[2] Tusch, R.; Kosch, H.; Boszormenyi, L. VIDEX: An Integrated Generic Video Indexing Approach; ACM Multimedia, 2000; pp 448-451

[3] Tran, D. A.; Hua, K. A.; Vu, K. VideoGraph:A Graphical Object-based Model for Representing and Querying Video Data; 2000; In Proc. of the 19th International Conference on Conceptual Modeling (ER2000), pp 383-396

[4] Combi, C. Modeling temporal aspects of visual and textual objects in multimedia databases; Int Workshop on Temporal Representation and Reasoning 2000; pp 59-86

[5] Hacid, M.-S.; Decleir, C. A database approach for modeling and querying video data. IEEE Transactions on Knowledge and Data Engineering 2000, 12, 729-750

[6] Yong, C.; De, X. Hierarchical semantic associative video model; In proc of IEEE International Conference on Neural Networks and Signal Processing, 2003; pp 1217-1220

[7] Aygun, R. S.; Yazici, A. Modeling and Management of Fuzzy Information in Multimedia Database Applications; Technical Report, 2002

[8] Arslan, U. A Semantic Data Model and Query Language for Video Databases; Master thesis, 2002

[9] Ahmet Ekin, Sports Video Processing for Description, Summarization, and Search (Chapter 2 Structural and Semantic Video Modeling), phd thesis, 2003

[10] Yu Wang, Chunxiao Xing, Lizhu Zhou, THVDM: A Data Model for Video Management in Digital Library, proceedings of the 6th International Conference of Asian Digital Libraries, 2003, pp 178-192

[11] Yu Wang, Lizhu Zhou, Chunxiao Xing. An Evaluation Method for Video Semantic Models, International Workshop on Multimedia Information Systems (MIS) 2005, pp207-220

[12] Yu Wang, Lizhu Zhou, Jianyong Wang. Model Video Semantics with Constraints Considering Temporal Structure and Typed Events, accepted by ICDE06 PHD workshop

[13] José M. Martínez. Mpeg 7 Overview, 2004 http://www.chiariglione.org/MPEG/standards/mpeg-7/mpeg-7.htm

[14] Chunxiao Xing, Lizhu Zhou et al. Developing Tsinghua University Architecture Digital Library for Chinese Architecture Study and University Education.ICADL2002,pp206-217

Using MILOS to Build a Multimedia Digital Library Application: The PhotoBook Experience[*]

Giuseppe Amato, Paolo Bolettieri, Franca Debole, Fabrizio Falchi, Fausto Rabitti, and Pasquale Savino

ISTI-CNR, Pisa, Italy
firstname.lastname@isti.cnr.it

Abstract. The digital library field is recently broadening its scope of applicability and it is also continuously adapting to the frequent changes occurring in the internet society. Accordingly, digital libraries are slightly moving from a controlled environment accessible only to professionals and domain-experts, to environments accessible to casual users that want to exploit the potentialities offered by the digital library technology. These new trends require, for instance, new search paradigms to be offered, new media content to be managed, and new description extraction techniques to be used.

Building digital library applications, and effectively adapting them to new emerging trends, requires to develop a platform that offers standard and powerful building blocks to support application developers. In this paper we discuss our experience of using MILOS, a multimedia content management system oriented to the construction of digital libraries, to build a demanding application dedicated to non-professional users. Specifically, we discuss the design and implementation of an on-line photo album (PhotoBook), which is a digital library application that allows people to manage their own photos, to share them with friends, and to make them publicly available and searchable.

PhotoBook, uses a complex internal metadata schema (MPEG-7) and allows users to simply express complex queries (combining similarity search and fielded search), enabling them to retrieve material of interest even if metadata are imprecise or missing.

1 Introduction

Many Digital Library Applications (DLA) supporting an effective and efficient archiving and retrieval of documents with different types of data and content, and that are used for many diverse purposes, have been developed. However, Digital Library technology will arrive at a definitive maturity when powerful and simple software development tools that enable the development of DLAs, will be available. Regrettably, nowadays several Digital Library Applications are monolithic software modules where the application itself, the content management software, and the digital library are merged

[*] This work was partially supported by the DELOS NoE, funded by the European Commission under FP6 (Sixth Framework Programme), by the ECD and VICE projects, both funded by the Italian Ministry of Research and Education. The PhotoBook Application can be accessed from http://milos.isti.cnr.it/

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 379–390, 2006.

together. Many systems were built having in mind a specific application and, in many cases, a specific document collection, thus resulting in an ad-hoc solution: all components of the DLA – the data repository, the metadata manager, the search and retrieval components, etc. – are specific to a given application and cannot be easily used in other environments. In these cases the development of the DLA is expensive and it is not possible to personalize or specialize the DL services and to adapt them to new user requirements.

An important building block of any DLA is the module responsible for the archiving, access, and retrieval of data objects. In many cases, as types of data managed change, or when different metadata format are used, or if different search functionality is required, a new component is build. Standard building components to help digital library application developers are needed.

The aim of this work is to illustrate the use of MILOS, a Multimedia Content Management System (MCMS) for digital libraries, to develop a demanding application: the PhotoBook Digital Library (http://milos.isti.cnr.it).

The advent of digital photography, combined to the wide access to internet resources, made popular the creation of personal and publicly accessible distributed photo archives. Typical examples are the *Flickr* service (www.flickr.com), the *snapfish* service offered by HP (www.snapfish.com), and *Picasa* (picasa.google.com). Users of these archives may create and manage their own photo albums, decide who and how can access their photos, provide a description of the photos to simplify their access. This simple yet powerful application, poses several complex requirements to the MCMS component, which must support: (i) the distributed storage and classification of photos, (ii) the description of photo content through an appropriate metadata model, (iii) the search based on photo description and photo content, (iv) the management of personal folders (photo albums), (v) the controlled access management. Some of the mentioned functionality, do require the development of an appropriate Web-based user interface. A final, but significant requirement is that the development of the application must be simple and fast. The tools previously mentioned (flickr, snapfish, and Picasa) offer most of these functionality but with search capabilities limited to manually associated metadata tags. Furthermore, as mentioned at the beginning of this paper, they are ad-hoc applications, whose development required significant investments and whose extensions would also require considerable efforts. Our aim is to show how, by using the general purpose MCMS system MILOS, is possible to build an application with similar functionality – and even with more powerful search capabilities – with a limited effort.

The development of the Photobook application has a twofold purpose: i) testing how the most recent digital library technology is able to deal with these new emerging trends, and ii) to verify if the MILOS multimedia content management system is capable of providing effective tools to rapidly develop a multimedia digital library application with high demanding requirements.

In the following we will briefly resume the functionality of the MILOS MCMS. We will discuss the functionality of the PhotoBook Digital Library Application. Then, we will present the architecture of PhotoBook and the specific functionality offered and the solutions adopted for picture management, metadata management – with particular attention to automatic metadata extraction – and photo search.

2 MILOS Overview

The MILOS system [4,1] is a Multimedia Content Management System with a number of characteristics that make it particularly suitable for the development of Digital Library applications. In particular, as we will underline in this section, MILOS functionality are particularly appealing for the development of the Photobook application, which requires the archiving of a large number of images from users distributed over the internet, efficient content-based retrieval of photos, according to their metadata values and based on their physical content, creation of a Web interface for the storage, query formulation and presentation of results.

Key characteristics of the MILOS system are the **flexibility** in managing different types of data and metadata and the independence from the specific format used to represent them. This implies that the application developer is not required to specify the details of the storage strategies used and the details of the access methods to be adopted; he/she only needs to specify the characteristics of the data and metadata and the functionality that are required, such as the requirement of storing high resolution photos, and supporting their efficient access based on the combination of metadata attributes and physical characteristics of photos. The flexibility of the MILOS system is also related to the possibility of developing end-user applications which are independent from the modality used to store data and search them. In particular, it is possible to store data described with a specific metadata model and to search them by using a different model.

Another key characteristic of MILOS is the **efficiency** in storing and searching multimedia objects. This requires a system which is (a) scalable when the size of the archive and the number of users accessing the application varies, and (b) efficient when processing complex queries on metadata values and data object's content.

The MILOS system has three main components: the Metadata Storage and Retrieval (MSR) component – dedicated to support the storage of metadata and the content-based retrieval of multimedia objects – the Multi Media Server (MMS) component – dedicated to support the storage and access of multimedia content – and the Repository Metadata Integrator (RMI) component – dedicated to provide the independence between the metadata format used by end-user applications and the internal metadata formats.

Let us provide some more details on the MSR, which is of particular importance in the PhotBook application. Details on the MMS and RMI are provided elsewhere [4].

2.1 Metadata Storage and Retrieval

The MILOS MSR, despite the approaches adopted by other similar systems, uses an enhanced native XML database/repository system with special features for DL applications. Indeed, XML represented metadata may have arbitrary complex structures, which allows one to deal with complex metadata schemas, and can be easily exported and imported.

The MILOS XML database [2] stores and retrieves any valid XML document. No metadata schema or XML schema definition is needed before inserting an XML document, except optional index definition for performance improvement. Once an arbitrary XML document has been inserted in the database it can be immediately retrieved. This allows DL applications to use arbitrary (XML encoded) metadata schemas and to deal with heterogeneous metadata, without schema design constraints and/or overhead due to metadata translation.

This special purpose native XML database/repository system is much simpler than a general purpose commercial XML database system, but with much better performances where needed. It supports standard XML query languages (XPath [6] and XQuery [7]) and it offers advanced search and indexing functionality on XML documents. The MILOS XML database supports high performance search and retrieval on heavily structured XML documents, relying on specific index structures [3,14], as well as full text search [13], automatic classification [8], and feature similarity search [15,5]. This is particularly relevant in the Photobook application, where metadata photos are represented in the MPEG-7 format: each photo includes in the description also features automatically extracted from visual properties, such as color histograms, textures, shapes, etc.

To deal easily and transparently with these advanced search and indexing functionalities, the syntax of the basic XQuery language was extended to deal with approximate match and ranking. Specifically the $\sim$ operator is used for approximate match (that is, feature similarity and text search). More details can be found in [2].

3 The PhotoBook Digital Library Application

The use of digital cameras is becoming very popular in the society. People can take picture and immediately see the result by downloading the images on their personal computers, avoiding the cost of printing them. Digital technology applied to the personal cameras has changed the way people use and manage their pictures. People take much more pictures, given that they can immediately see them at no cost, print a significant minority of them, and especially share them with friends by sending them by email, or more frequently by storing and publishing them on on-line photo albums.

Several sites that offer popular on-line photo albums services are nowadays available on the internet and they are continuously enriched by new services to help people on one hand to publish and share their material, on the other to access and search for published material.

Even if on-line photo albums are mainly addressed to non-professional users, they can be reasonably considered as digital library applications. In fact, such systems typically offer important functionality also offered by professional digital library systems. As an example, on-line photo albums offer storage and preservation services, privacy and rights management, personalization, metadata editing and annotation functionality, advanced search functionality.

Two distinct types of users can be broadly identified for such systems. Users can act as *publishers* when they publish, annotate, classify, and manage digital photos. Users can act as *searchers* when they search for published digital material. Both publishers, and searchers are non-professional users of these on-line services. The two roles sometime can be merged. Many people are in fact at the same time publishers and searchers.

On-line photo albums nowadays represent an emergent phenomenon of the internet society. Several stand-alone on-line photo album services are in fact available on the internet. In addition, almost all internet providers, portals, search engines, photo camera producer, and photo printing services have added an on-line photo albums to the services that they already provide.

In the following we discuss the requirements of the PhotoBook (http://milos.isti.cnr.it/) application , its architecture and the various components.

3.1 PhotoBook Application Functionality

The PhotoBook application is intended to be accessible on-line by everybody. It offers services to two categories of users: *non-registered users* and *registered users*.

Non-registered users can use the PhotoBook application to search for and access photos that were uploaded in the system and made publicly available.

Registered users in addition to the capabilities of the non-registered users can also

1. upload photos in the system
2. manage their own photos
3. share their photos with friends
4. make their photo publicly available

Each photo managed by the PhotoBook has one single owner, which is the user that uploaded it (we suppose that a person that uploads a picture owns it or he/she has the right to do it). The owner can set a photo to be *private* or *public*.

All users (registered and non-registered) should be allowed to search and access digital pictures that were uploaded in the system and that were marked by the owner as public.

Registered users can organize their own photos in *albums*. An album is basically a *collection* of photos. For instance, a user can decide to create an album containing the photos that he/she took during last summer. A photo may belong to several albums at the same time. Users can share an album (containing both private and public photos) with friends. A private photo can be accessed just by the owner of the photo and by his friends, when they are given an handler to an album that contains it. A user's friend does not need to be registered in the system. Access to an album is obtained by using a system-generated URL containing the handler to the album. Every non-registered user can access an album using the album handler.

The owner of an album can remove an album at any time. An album deletion does not delete the photos that it contains, which must be explicitly removed from the system if needed.

When a registered user uploads a new photo, he/she can associate it with some descriptions related to the picture's content. Descriptions can be changed and refined afterward. A bulk upload functionality is also supported to easily insert photos having common descriptions. For instance, a user can insert, with a single action, all photos taken during last weekend in the mountains. In this case, all inserted photos are associated with the same common description.

In addition to descriptions created by the users, the system automatically analyzes photos to extract additional metadata. Specifically, feature descriptors that enable similarity search are automatically extracted. Similarity search [15] allows users to search for pictures similar to pictures chosen as queries. This possibility can be particularly useful to retrieve poorly described pictures. Consider that, as previously stated, the PhotoBook application is addressed to a non professional target. Therefore, imprecise, erroneous, incomplete, or completely missing descriptions may frequently occur. Similarity search is an option for searching for photos of interest, which is really useful especially in this non-professional context. With similarity search, a user can be able to retrieve, for instance, pictures of the tour Eiffel by using another picture of the tour Eiffel as a query, even if the retrieved pictures were not correctly annotated by their owner.

Metadata manually and automatically generated are represented by using standard metadata schema. Specifically we have used MPEG-7 [9] metadata schema, given that it is able to represent both low level feature descriptors, for similarity search, and more conceptual descriptions. However, given that the PhotoBook application is intended to be used by non professional users, the metadata complexity is hidden to them. This implies that information is presented to the user in a natural and intuitive way and that users are asked to insert a minimal amount of descriptions when uploading photos. A relevant part of information is obtained automatically, and the system is able to satisfy most of the user requirements even with incomplete, erroneous, or partial information. For instance, in addition to image analysis to support similarity search, a lot of more conceptual information can be obtained exploiting file names and folders names.

During photo management and photo searching activities, users see on the screen the photo thumbnails. However, a user can download full size pictures, if needed.

3.2 PhotoBook Architecture

The architecture of the PhotoBook application is sketched in Figure 1. It is a classical n-tier application, where the data layer, the application layer, the presentation layer, and the client layer are responsible of different classes of activity of the application.

At the data layer we have MILOS, which is responsible for the management, access, and retrieval, of all types of data of the application. Data managed by MILOS include pictures uploaded by users, metadata, and user data.

At the application layer we have a set of components responsible of mediating the interaction of the modules that implement the web user interface with MILOS. These components actually implement the application's logic, separately and independently from the user interface logic and the data management logic. These components are responsible of checking the credentials of users, organizing retrieved result on behalf of the application logic, translating the user queries into correct queries to MILOS, building valid metadata on the basis of the user input, implementing storage strategies for the uploaded material (relying on the MILOS features), and generating additional metadata by analyzing uploaded data. Interaction with MILOS is obtained using the SOAP protocol, being MILOS a Web service. These components are also implemented as web services.

The presentation layer contains the software modules in charge of drawing the user interface on behalf of the client layer. PhotoBook is a web application: accordingly, the interface layer is composed of Java Server Page (JSP) modules which dynamically generate the web user interface. These modules interact with the application layer by using SOAP.

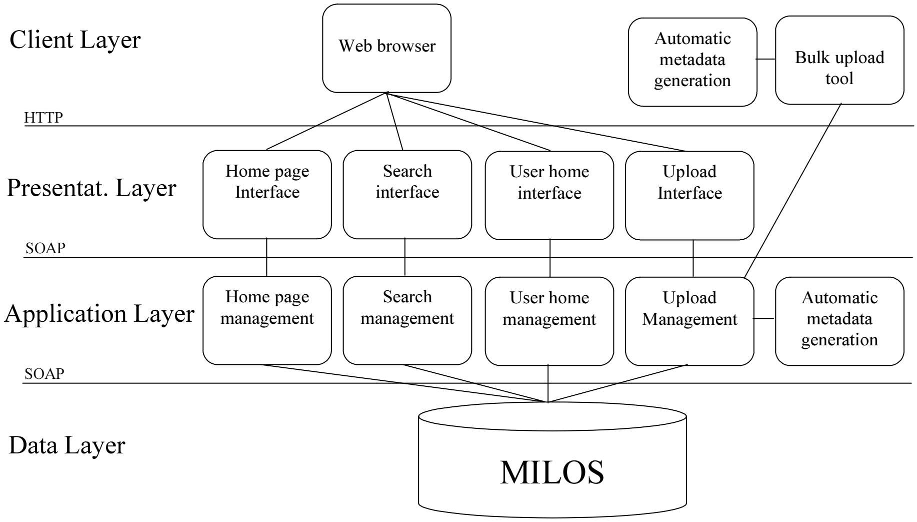


Fig. 1. Architecture of the PhotoBook Application

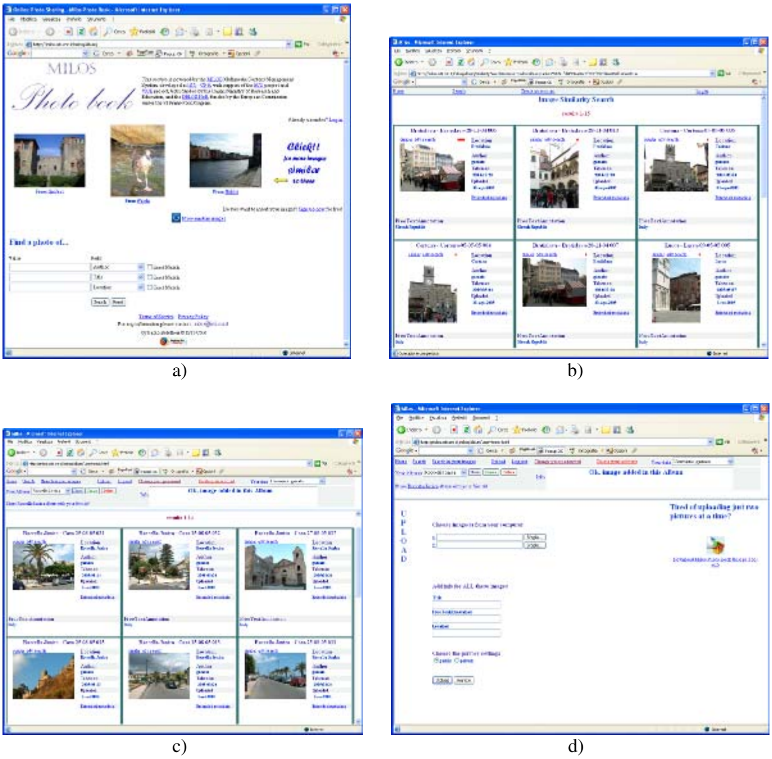
Clients access PhotoBook by using a normal web browser. In addition at the client side, users may optionally use a Java bulk-load component that allows them to upload several pictures in the system, and to locally analyze them to automatically generate additional metadata. The use of the bulk-load tool, moves part of the logic for metadata generation to the client side, distributing the burden of image analysis among several clients and improving upload performance of the system.

3.3 The PhotoBook User Interface

A sketch of the main web interfaces is shown in Figure 2. The home page of the PhotoBook application is shown in Figure 2a. From the home page, every user (registered and non-registered) can search for public material on the system, login for managing the owned material, registering into the system. From the home page users can search for pictures by using a fielded search or similarity search. In case of fielded search users can search for pictures by expressing restrictions on the owner of the pictures, the location where they were taken, their title, and on the textual description of the pictures. In case of similarity search, the user can search by choosing a picture among those randomly proposed by the system. Random pictures can be renewed on demand by the user.

Results are shown in the search page (Figure 2b). From there, users can refine their queries by choosing a picture in the result to submit a new similarity search or to submit a complex search query, which combines similarity and fielded search. For instance, a user can search for images similar to the chosen one, whose location is Pisa.

Registered users that enter the system (Figure 2c), can also manage their own collection of pictures. Registered users can update metadata associated with pictures, organize pictures in thematic albums, share albums with friends, and upload new pictures.



a)

b)

c)

d)

Fig. 2. The PhotoBook interface (http://milos.isti.cnr.it). a) The home page of the PhotoBook application, b) The search interface of the PhotoBook application, c) The registered user page of the PhotoBook application, d) The upload page of the PhotoBook application.

There are two options for uploading new pictures. Registered users can use the web interface (Figure 2d) or a bulk-load tool running on their computers. By using the web interface users can upload up to two pictures at time, by specifying common descriptions for both. In this case, automatic metadata extraction for similarity search is executed on the PhotoBook server. By using the bulk-load tool, users can submit an arbitrary amount of pictures from their hard disk. In this case, users can assign common descriptions to the uploaded picture. In addition the bulk-load tool can refine these description by using information extracted from the folder names and file names. Image analysis for similarity search is also performed on the user computer. By using the upload tool, the PhotoBook server has just to register the insertions, store the pictures, and update the internal indexes.

4 Picture Management

MILOS allows different storage strategies to be used when dealing with different types of media. The use of specific storage strategies is transparent to the application. To obtain that, MILOS (the MMS component) identifies all documents with an URN and maintains a mapping table to associate URN with actual storage locations and access protocols. When an application requires to retrieve a document, it uses the URN to make the request to the MMS, which returns the URL to be used to access the document.

In the PhotoBook application we decided to store pictures in a normal file system, and to access them via a web server using the HTTP protocol. Pictures belonging to a user are stored in a personal folder. Two different versions of every pictures are stored: a thumbnail version and a full size version. Thumbnails and full version pictures are stored in two different subfolder of the user's folder.

An example of URN used in the PhotoBook is the following:

urn:milos:album:Paolo:image_jpeg:7897056ecc55cd6f7cffa78413e4e2ac

which refers to a jpeg image of the user Paolo. When the PhotoBook application wants to retrieve a picture, it uses the getURL(URN,version) method of MILOS to obtain the URL to be used to retrieve it. When a request for the full size version of previous URN is receioved, MILOS returns the following URL:

http://milos.isti.cnr.it/milos-MMS/MMS/album/ Paolo/big/image_jpeg/
/image_jpeg7897056ecc55cd6f7cffa78413e4e2ac.jpg

5 Metadata Management

MILOS is able to deal with any metadata schema. The only requirement is that metadata should be encoded in XML. No schema definition is needed to instruct MILOS to deal with a specific metadata schema. Once an XML document is inserted in MILOS, it is immediately available to be searched and accessed, using the advanced search functionality of the native XML search engine embedded in MILOS.

This feature gave us a lot of freedom and flexibility in the choice of the metadata to be used for the application and freed us from the difficulties of instructing a specific search engine or database to deal with them.

We decided to use MPEG-7 metadata [9] for the image description. Figure 3a gives an example of the MPEG-7 description of a picture. An MPEG-7 description contains low level features to be used for similarity search, conceptual content descriptions, usage rights, creation time information, etc. Specifically, the `<VisualDescriptor>` tags, in the figure, contain scalable color, color layout, color structure, edge histogram, homogeneous texture information to be used for image similarity search. MILOS indexes this tag with a special index to offer efficient similarity search. Conceptual information is contained in the `<Title>`, `<Abstract>`, and `<Location>` tags. Right management information is contained in tags `<UsageInformation>`. Creation time information is maintained in `<CreationTool>` tag. Specifically, eXtended Image File Format (XIFF) information is automatically extracted from pictures. This information includes, for instance, parameters set in the camera when the photo was taken, the type of camera, etc. The MPEG-7 standard prescribes that metadata must be encoded in XML. Therefore, its use was straightforward in MILOS.

Users can create thematic albums containing pictures related for instance to an event. Albums are also encoded in XML as shown in Figure 3b. An album has basically a name, an owner, and a set of pictures included in the album.

```
<Mpeg7 schemaLocation="urn:mpeg:mpeg7:schema:2001 Mpeg7-2001.xsd" >
  <Description type="ContentEntityType" >
   <MultimediaContent type="ImageType" >
    <Image>
     <MediaLocator>
        <MediaUri>
urn:milos:album:paolo:image_jpeg:7897056ecc55cd6f7cffa78413e4e2ac
        </MediaUri>
     </MediaLocator>
     <VisualDescriptor type="...." .... >....</VisualDescriptor>
     <UsageInformation>
             <Rights> <RightsID>paolo:public</RightsID> </Rights>
       </UsageInformation>
     <CreationInformation>
      <Creation>
       <Title>Tour in Umbria</Title>
       <Abstract><FreeTextAnnotation>...</FreeTextAnnotation></Abstract>
       <Creator><Role href="creatorCS" ><Name>creator</Name></Role>
          <Agent type="PersonType" >
             <Name><GivenName>paolo</GivenName></Name>
          </Agent>
       </Creator>
       <CreationCoordinates>
        <Location><Name>Perugia</Name><Region></Region></Location>
       </CreationCoordinates>
       <CreationTool>
                    ....eXtended Image File Format (Xiff) data....
        <Tool><Name>Canon Canon PowerShot A85</Name></Tool>
       </CreationTool>
      </Creation>
     </CreationInformation>
    </Image>
   </MultimediaContent>
   <DescriptionMetadata>
    <CreationTime>Aug 5, 2005</CreationTime>
   </DescriptionMetadata>
  </Description>
 </Mpeg7>
```

a)

```
<album>
 <owner>gamato</owner>
 <name>Roccella Jonica</name>
 <id>fead7ff5333fa069401abe8bf2521d3d</id>
 <urn>.....</urn>
 <urn>.....</urn>
 .....
</album>
```

b)

```
<login>
   <username>gamato</username>
   <password>sjadfe</password>
   <email>gamato@interfree.it</email>
...
</login>
```

c)

Fig. 3. Metadata used in the PhotoBook application: a) Mpeg7 metadata to describe the pictures, b) representation of albums (thematic collection pictures), c) user data

User data are also very simple. Figure 3c shows how they are represented in PhotoBook. A user has a username, a password, and an email address.

5.1 Automatic Image Processing

Feature extraction was performed employing an application we built upon the MPEG-7 experimentation model (XM, [11]) of MPEG-7 Part 6: Reference Software. The software can extract all MPEG-7 image Visual Descriptors defined in [10]. For the PhotoBook we extract 5 MPEG-7 descriptors: ScalableColor (a color Histogram in the HSV Color Space), ColorStructure (captures both color content and information about the spatial arrangement of the colors), ColorLayout (represents the spatial layout of color images), EdgeHistogram (spatial distribution of five types of edges), and HomogeneousTexture (characterizes the properties of texture in an image). For all thoose descriptors the suggested distance functions [12] are metric. The result of the extraction process is an XML document like the one in Figure 3 without usage and creation information. The values inside the `<VisualDescriptor>` tags are integer vectors (ScalavbleColor, ColorSTructure and EdgeHistogram) or more complicated XML subtrees with integers as values (ColorLayout and HomogeneousTexture).

6 Search Capabilities

The MILOS native XML database/repository supports high performance search and retrieval on heavily structured XML documents, relying on specific index structures

[3,14], as well as full text search [13], automatic classification [8], and feature similarity search [5]. This is compatible with current trends of the new generation of XML encoded metadata standards, such as MPEG-7, which include in their description also features automatically extracted from visual documents, such as color histograms, textures, shapes, etc. Specifically, the MILOS XML database allows the system administrator to associate specific XML element names with special indexes. Therefore, for instance, the tag name `<abstract>` can be associated with a full text index. On the other hand, the MPEG-7 `<VisualDescriptor>` tag can be associated with a similarity search index structure.

These features of the MILOS system were very useful to provide users of PhotoBook with advanced search functionalities and to provide developers with all needed functionality for picture description, album, and user management.

Specifically the PhotoBook application allows users to use similarity search, by exploiting the `<VisualDescriptor>` tags included in the MPEG-7 metadata. Users can also submit full-text queries, by using the full-text descriptions included in the `<abstract>` tags. In addition, users can perform a search by expressing queries that use any tag content of the MPEG-7 metadata.

Users can also express complex queries, where full-text, fielded, and similarity search is conveniently combined.

7 Conclusions

This paper illustrates the main characteristics, architecture and design choices adopted in the PhotoBook Digital Library application, which supports archiving, indexing, sharing and content-based search of photos. PhotoBook was built by using a general purpose Multimedia Content Management System, MILOS, which is specifically designed to create high performance Digital Library applications. By using MILOS we had several advantages: (a) the development of the entire application was realized with a limited effort – approximately one month of work of an experienced programmer, that developed the user interface of the application, (b) powerful and efficient content-based search capabilities have been included, (c) flexible storage management is possible – for example, if the size of the archive will increase in the future, the storage strategies can change without any modification to the application and transparently for end-users, (d) integration with other similar archives, based on different metadata formats, or archiving of photos represented in metadata formats different from that used in PhotoBook, can be easily obtained.

Future research efforts will be spent in extending some MILOS functionality: the development of the PhotoBook application has shown that tools to be used for the development of the application user interface would greatly reduce application development costs; on the other side, by improving the automatic semantic analysis of multimedia content, it should be possible to support more powerful content-based searches.

The extended MILOS functionality will then be used to provide a new version of the PhotoBook application, with the extended classification, indexing and search capabilities. Finally, it is our intention to continue the test of the MILOS system by developing other Digital Library applications.

References

1. MILOS: Multimedia dIgital Library for On-line Search. http://milos.isti.cnr.it/.
2. G. Amato and F. Debole. A native xml database supporting approximate match search. In *Europeean Conference on Digital Libraries, ECDL 2005, Vienna, AT, September 18-23 2005*, 2005.
3. G. Amato, F. Debole, F. Rabitti, and P. Zezula. YAPI: Yet another path index for XML searching. In *ECDL 2003, 7th European Conference on Research and Advanced Technology for Digital Libraries, Trondheim, Norway, August 17-22, 2003*, 2003.
4. G. Amato, C. Gennaro, F. Rabitti, and P. Savino. Milos: A multimedia content management system for digital library applications. In *Europeean Conference on Digital Libraries, ECDL 2004, Bath, UK, September 12-17 2004*, 2004.
5. C. Böhm, S. Berchtold, and D. Keim. Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases. *ACM Computing Surveys*, 33(3):322–373, September 2001.
6. W. W. W. Consortium. XML path language (XPath), version 1.0, W3C. Recommendation, November 1999.
7. W. W. W. Consortium. XQuery 1.0: An XML query language. W3C Working Draft, November 2002. http://www.w3.org/TR/xquery.
8. N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambidge University Press, 2000.
9. ISO/IEC. Information technology - Multimedia content description interfaces. 15938.
10. ISO/IEC. Information technology - Multimedia content description interfaces. Part 3: Visual. 15938-3:2002.
11. ISO/IEC. Information technology - Multimedia content description interfaces. Part 6: Reference Software. 15938-6:2003.
12. P. Salembier and T. Sikora. *Introduction to MPEG-7: Multimedia Content Description Interface*. John Wiley & Sons, Inc., New York, NY, USA, 2002.
13. G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company, 1983.
14. P. Zezula, G. Amato, F. Debole, and F. Rabitti. Tree signatures for xml querying and navigation. In *Database and XML Technologies, First International XML Database Symposium, XSym 2003*, volume 2824 of *LNCS*, pages 149–163. Springer, 2003.
15. P. Zezula, G. Amato, V. Dohnal, and M. Batko. *Similarity Search - The Metric Space Approach*, volume 32 of *Advances in Database Systems*. Springer, 2006.

An Exploration of Space-Time Constraints on Contextual Information in Image-Based Testing Interfaces

Unmil Karadkar[1], Marlo Nordt[1], Richard Furuta[1], Cody Lee[2], and Christopher Quick[2]

[1] Center for the Study of Digital Libraries, College Station, TX 77843-3112, USA
{unmil, mnordt, furuta}@csdl.tamu.edu
[2] Michael E. DeBakey Institute, College Station, TX 77843-4466, USA
{codylee, cquick}@tamu.edu

Abstract. Digital image collection interface layouts vary in the nature and degree of contextual information they provide to their users, thus enabling or impeding specific tasks. We are exploring image presentation techniques to support image-centric cognitive tasks in the context of cardiovascular systems research and education. To investigate the effect of image layout on user performance, we conducted an experimental evaluation of three image layouts for three representative tasks in this domain. The layouts varied the spatial and temporal presentation of images, thus providing different contextual information to the test subjects. Our results indicate that the degree of contextual information provided by the image layouts affected user performance, as did their research expertise. These results will inform the design of user interfaces for performing image-focused cognitive tasks as well as the development of interfaces for training novice researchers.

1 Introduction

Government agencies, educational institutions, scientific communities, commercial entities, and individuals all develop and deploy Web-based digital image collections to further their goals. The two-dimensional, grid-based, thumbnail layout has emerged as the de-facto standard for browsing Web-based image collections and search results [1, 6, 12, 19]. While the Frappr! interface employs a novel map-based image overlay, it retains the familiar 2-D thumbnail layout for displaying the images associated with each location [10].

A significant body of research has focused on designing targeted interfaces for presenting digital images in the context of personal photograph collections and digital libraries. These interfaces employ a variety of techniques, such as zooming [4], radial quantum layouts [18], collaging [13, 15], temporal cues [11, 14], 3-D immersive environments [5], and mechanisms for creating topical sub-collections [17, 20] to support users in recognition tasks such as searching and browsing large image collections for finding images that match specific criteria.

In contrast, to the recognition-oriented tasks, researchers and technicians in medical or atmospheric science areas evaluate images for identifying patterns and for answering specific questions. For example, doctors use X-ray images to determine the possibility and extent of bone fractures and weather forecasters predict the arrival of

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 391 – 402, 2006.

storm fronts. In these domains, trained professionals analyze images and apply their knowledge from the real world to reach critical conclusions.

We are exploring techniques for testing specific research skills and for training novice researchers in methods of cardiovascular research using a video and image corpus. In particular, we are investigating the role of image layouts and the contextual information embodied by these layouts in image-based question-answering tasks. Towards this purpose, we conducted an experimental evaluation of three image layouts that differed in the spatial and temporal organization of images. The layouts we chose varied the nature of the contextual information available to test subjects as well as the amount of this information. Our subjects performed three tasks with each of these layouts to test for specific, representative skills that cardiovascular systems researchers acquire through experience.

Our results indicate that the degree of contextual information provided by the interfaces affected user performance, as did their research expertise. The results highlight significant differences in the strategies used by experts and novices and will inform the design of software for training cardiovascular systems researchers. These results have broader applications in areas that involve image-based cognitive tasks.

The remainder of this paper is organized as follows: the next section situates our research in the context of the cardiovascular systems research domain. We then describe the experimental design of the evaluation: the nuances of the three image layouts and the three representative tasks that evaluate cardiovascular systems research skills, followed by the findings of our study. We discuss the implications of these experimental results and conclude by presenting continuing research directions.

2 Cardiovascular Research

Researchers in the Cardiovascular Systems Dynamics Laboratory, also known as the “bat lab”, conduct cardiovascular research through non-invasive, *in-vivo* studies on Pallid bats. The bat lab houses a colony of bats and extensive equipment for the study of their blood and blood vessels including high-resolution microscopes connected to computers for recording the experiments in a high-quality video format. The Pallid bat’s thin and transparent wing enables researchers to inspect blood cells, vessel walls, and much of the cardiovascular network visually through a microscope. While bats do not, by nature, stick their wings underneath microscope objective lenses, they have been trained to sleep in a special container designed for this purpose with one wing extended. Researchers in the bat lab observe the wings of napping bats and often save video of the microscope feed of these experiments for later analysis.

2.1 Research Skills

Typically, researchers learn basic concepts by viewing microscope video feeds of the experiments where they can observe the bat’s cardiovascular system. They view a bat wing in its natural condition, as well as by modifying these conditions, such as by applying pressure to occlude blood-flow to a specific part of the wing. Researchers can identify features such as major and minor arteries and veins, capillaries, and lymphatic vessels. Some features, such as blood vessels, are easy to identify as well as classify. Experienced individuals can quickly gauge whether they are looking at a vein, a second-level venule, or a capillary. Other features, such as lymphatic vessels, are more difficult to recognize. Researchers often use strategies that employ external knowledge in order to locate these features. For example, following a vein longitudinally is likely to help locate physical proximate lymphatic vessels. Acquisition of such skills helps them minimize the time required for basic tasks such as identification of features and focus on the research questions regarding these features.

2.2 Image-Based Research Skills Assessment

Our observation, that while conducting experiments, researchers identify basic features on the bat wing almost instantaneously, led us to explore the possibility of using still images instead of videos for testing these skills. Unlike videos, several images can be displayed simultaneously, and in different layouts, thus supporting differences in users' mental models and practices. We tested the use of images for specific tasks with the aid of a pilot user—an advanced graduate student—who effortlessly identified various features on a bat wing using printed photographs. Encouraged by the performance of this user, we designed an experimental evaluation to study the use of still image interfaces for testing basic cardiovascular research skills.

3 Evaluation of Image Layouts for Cardiovascular Research

The evaluation targeted two goals: identification of image layouts for testing critical skills and exploration of digital image collections for training novice researchers. We designed a set of experimental tasks around basic skills that cardiovascular systems researchers apply routinely in the course of their experiments. Performing these tasks involved the use of information gained from neighboring images in a layout. To investigate the role of available contextual information, we chose three image presentation layouts that provide different contextual information to their users.

3.1 Subject Characteristics

Our subject pool comprised of 15 graduate and undergraduate student researchers in the bat lab with varying degree of experience with analyzing bat wing videos. These students, 11 male and 4 female, were between 18 and 34 years old. As active bat cardiovascular researchers, the subjects possessed adequate knowledge of the area for participating in this evaluation without additional training. We classified the subjects who had worked with bat videos for one complete semester in the lab as novices (6 subjects) and those who had completed two or more semesters as experts (9 subjects).

3.2 Image Layouts

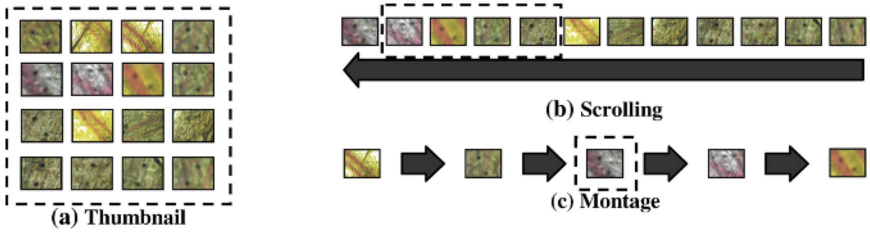
The subjects worked with the three different image layouts diagrammed in figure 1. These layouts provided users with diverse contextual information due to differences in spatial and temporal organization of images illustrated in table 1. The images enclosed by a dashed line in figure 1 correspond to the last column in table 1 and were displayed simultaneously to the users. All layouts displayed individual images of the same size in order to ensure that the size of the images, and hence, the level of visible detail did not affect subjects' perception of the interfaces. We created these layouts by placing the images in a PowerPoint presentation.

Table 1. Contextual differences in image layouts

Layout	Spatial Dimensions	Temporal Dimension	Concurrent Images
Thumbnail	2	No	16
Scrolling	1	Yes	4
Montage	0	Yes	1

Thumbnail. The thumbnail layout displayed the images in a 4x4 grid, illustrated in figure 1(a), which allowed us to fit 16 images on one screen while keeping the individual images at a workable size. This layout is widely used for presenting image search results [1, 6, 12] and as a browsing interface for Web-based digital image collections [19, 24]. Simultaneous display of all images enables users to contextualize each image in terms of all others. Users can compare images and detect patterns within the sets. The thumbnail layout presents images in two spatial dimensions and does not employ the temporal dimension, as images in the set are static.

Scrolling Filmstrip. Images in the scrolling filmstrip layout, shown in figure 1(b), scrolled smoothly from the right to the left at a constant rate. Our subjects viewed these images from the left to the right and could not control the scrolling rate. The images were displayed at the same size as the thumbnail layout but only four images were visible at any one time. Thus, this layout restricted our subjects' contextual space. When working with an image, subjects could compare this image to its predecessors as well as successors thus providing a forward and backward context for the images. This layout used one spatial dimension, as the images scrolled horizontally, and the temporal dimension, as the displayed images changed over time.


(a) Thumbnail

(b) Scrolling

(c) Montage

Fig. 1. Testing interfaces – images enclosed by the dashed line were displayed simultaneously

Montage. The montage layout resembles a slide show where successive images replace earlier ones. This layout, shown in figure 1(c), displayed all the images for an equal amount of time and did not allow users to control this time. The direct context of each image in this layout consisted of a preceding single image, although subjects could have worked with a larger context by recalling additional preceding images. The montage layout did not support a preview or look-ahead mechanism to include the succeeding images in constructing image contexts. This layout used the temporal dimension but no spatial dimensions; all images were displayed at the same location.

3.3 Tasks

The subjects answered image-based questions regarding three of the research skills that they acquire via experiments on bat cardiovascular systems. We showed them sets of sixteen images via PowerPoint slides on a computer display. The set size was selected to use the computer's available display space optimally while retaining the integrity of the features presented within the images. The pilot subject verified that he could work with these images and layouts effectively. Our subjects performed the following three tasks, each of which emphasizes a different research skill and embodies a different thought process.

Artery/Vein Recognition. The subjects viewed images of bat blood vessels and responded whether the blood vessel indicated by an arrow pointing to it was an artery or a vein. For this task, the subjects used their prior experience and knowledge regarding the visual appearance as well as behavior of arteries and veins to recognize the nature of the indicated blood vessel. For example, the subjects were aware that arteries and veins usually run in pairs and veins are wider than arteries. In addition, while the blood in the arteries flows at a constant rate, the blood flow in veins pulses in a stop-and-go fashion. Thus, a picture in which the cells in a blood vessel are clearly visible is more likely to be that of a vein. We recorded the subjects' verbal responses for later analysis.

Size Estimation. The subjects estimated the diameter of the indicated blood vessel at the location of the arrow. This task required them to couple their expertise in identifying the class of a blood vessel (artery, second-level arteriole, or a capillary) with their knowledge of the size for this class of vessels to estimate the diameter of the indicated vessel as accurately as possible. Like the Artery/Vein recognition task, we recorded the subjects' verbal responses for later analysis.

Lymphatic Vessel Wall Identification. Lymphatic vessels blend with the surrounding tissue and, hence, are difficult to identify. To make this task tractable, we showed the subjects different images as the vessel went through a complete expansion-contraction cycle. The subjects then marked the location of the wall using PowerPoint's on-screen marker. This task required the subjects to compare the images within each set and analyze the differences between these images. This analysis, coupled with their knowledge of lymphatic vessels enabled them to identify the vessel boundaries, a feature that is generally unidentifiable in a single still image.

3.4 Image Data Set

We obtained about 300 digital images for this study from video feeds of past experiments. The Artery/Vein recognition and Size estimation tasks employed about half of these pictures. The images used for these tasks were taken at two magnifications: 10X and 40X. Some images were reused, but each image was used only once for each of the tasks. For the Lymphatic vessel identification task, we identified video sections that displayed a complete expansion-contraction cycle for the vessel and captured a set of 16 images equally spaced over this section.

We obtained definitive answers for each question that our subjects would answer. A panel of three researchers provided the answers for the Artery/Vein recognition task. In order to ensure that we had conclusive answers for the size estimation task, we calculated the exact diameter of the arteries and veins at the locations where our subjects would estimate these values using calipers. We noted the exact boundaries of the lymphatic vessels from pre-recorded videos of observations of these vessels. Using this information, we prepared a comprehensive key that served as a standard for evaluating the test subjects' answers.

3.5 Experimental Design

The subjects performed the three tasks sequentially, completing one task before they started another. We reordered the tasks across subjects in order to balance the experiments. A third of the subjects performed the Artery/Vein recognition task (A), followed by the Size estimation task (E), and finally, the Lymphatic vessel identification task (L). The other two sets of subjects performed the tasks in the order E, L, A, and L, A, E.

For each task, subjects viewed the different image layouts in sequential cycles, balanced across the tasks. For task A, subjects viewed the Thumbnail (T), followed by the Scrolling filmstrip (S), and finally, the Montage (M). For task E, they viewed the layouts in the order M, S, T, and for task L, in the order S, T, M. The subjects viewed two cycles of these layouts for the Artery/Vein recognition and Size estimation tasks and answered one question per image (16 questions per layout). For the Lymphatic vessel identification task, they only answered one question per layout as they were looking at a sequence of images. For this task, we had the subject view three layout cycles in order to compensate for the reduction in answers. In order to enable the subjects to work on the images out of order if desired, each image within a view was labeled by a letter.

The subjects had a fixed amount of time for each task regardless of image layout. We timed our tasks based upon the performance of our pilot user. For the Artery/Vein recognition task, the subjects had three seconds for each image (48 seconds for each view), for the Size estimation task, they had four seconds for each image (64 seconds for each view), and for the Lymphatic vessel identification task, they had two seconds for each image (32 seconds for each view). The evaluation time for each subject was about 30 minutes including the time required for orientation and debriefing. At the outset, we introduced the subjects to each of the layouts. At the beginning of each tasks, we illustrated the task via an example. Users also could perform a short practice task to acquaint themselves with the task more thoroughly, prior to performing the actual task.

4 Results

We collected data from the subjects through a demographic questionnaire, a task questionnaire, responses to the task-specific questions and observation of subjects and their strategies as they performed the tasks. All but two of our subjects (an expert and a novice) had worked with bat videos more than once a week over the course of their cardiovascular research experience. This section presents the results of ANOVA analyses of subject responses.

4.1 Artery/Vein Recognition Task

Our subjects reinforced the conventional wisdom that Artery/Vein recognition is one of the first skills that new cardiovascular researchers acquire by returning the best scores for this task. In terms of the different image layouts the subjects, experts as well as novices, consistently performed the best with the familiar thumbnail layout. The experts correctly answered 13.59 questions out of 16 on average (84.94%) to outperform the novices (p=0.005), who answered 9.75 questions correctly (61%). Figure 2 illustrates the differences in results of the two populations.

Overall, the experts' performance differed significantly across the individual layouts (p=0.005). They performed consistently on the two rounds of the familiar thumbnail layout (89.6%). On the other two layouts, they performed slightly worse than the thumbnail layout on the first round of both the montage (79.2%) and the scrolling layouts (75%), but their performance improved significantly (p=0.0004) on the second round as they quickly learnt the nuances of the new layouts. In fact, the experts performed better on the second round of the scrolling layout (91.4%) than they did on either of the thumbnail rounds. This suggests that once they adjusted to the "sense of impending doom"—as a user characterized the scrolling view—the scrolling and montage views may be better than the thumbnail view for supporting this task. Although the experts did well with the scrolling layout, most subjects (11 out of 15) thought that they were most effective on the thumbnail layout.

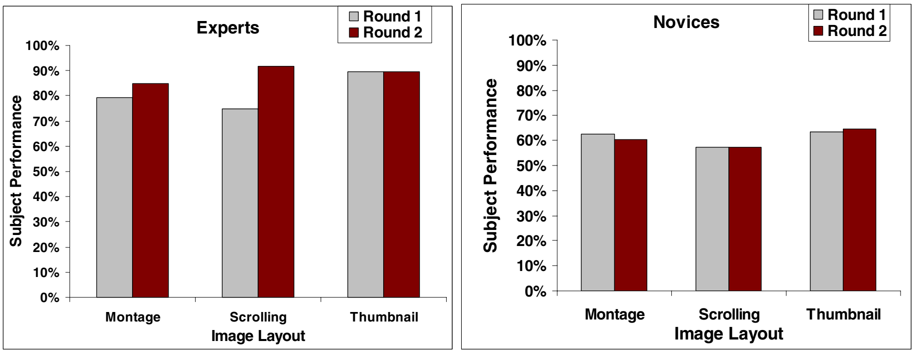


Fig. 2. Performance of subjects on the Artery/Vein recognition task

The performance of our novice subjects was consistent for each layout over the two rounds. While the experts improved with practice on the unfamiliar image layouts, we did not see any improvement with the novices. We believe that this is partly due to the difference in expertise of the two groups. While the experts were constrained by the layout and got better with practice, the novices were constrained by their experience rather than the presentation of the images or the contextual information provided by these layouts. Overall, the scrolling layout was the least favored, with only two subjects vouching for this layout.

4.2 Size Estimation Task

The subjects found it more difficult to estimate the diameter of arteries and veins than we had expected. While they regularly intuit the size of such features in their daily research in order to find features of interest, this general sense of proportion did not translate into accurate estimates. The subjects responded with a single number that was their best estimate for the diameter of the vessel in microns. As shown in figure 3, we graded their performance on this task at several levels. At the ±10% level, answers that fell within a 10% range of the exact answers were considered correct. Thus, for a vessel 40 microns in diameter, subject who guessed the vessel to be between 36 and 44 microns received credit for this question. At the ±40% approximation range answers between 24 and 56 microns were considered correct.

While the novices found it particularly difficult to estimate the sizes, even the experts could estimate only about 50% of the sizes within a ±40% range of the actual size. We did not find a significant difference in performance across the various image layouts. However, while the experts' responses on the scrolling and montage layouts were consistent between the two rounds, their performance on the thumbnail layout dramatically dropped for the second round. Overall, the experts significantly outperformed the novices on this task as well at all ranges of approximation. 56.5% of the experts' responses were acceptable at the ±40% range, compared to 25.3% of the novices' responses ($p<0.002$). At the ±10% range the acceptable responses were 13.2% and 4.5% respectively ($p<0.002$).

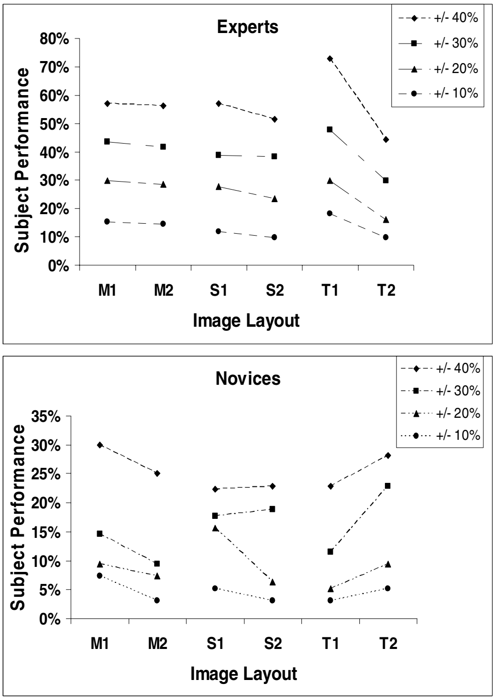


Fig. 3. Result of the Size estimation task

Some subjects liked the thumbnail view because they could compare the diameters of vessels across images. Thus, they estimated the diameter for a few images and compared the diameters across images to guess the others. We believe that this strategy may have backfired as the subjects may have missed the fact that the images were taken at different magnification factors, a fact that was stressed during their introduction to the task. Some novices had trouble managing the available time when working with the Thumbnail layout and did not have the time to attempt several questions. The subjects used the visual cues provided by the other layouts to respond to the images within the available time. 11 of the 15 subjects liked the thumbnail layout and although we did not find that the layout affected their performance, 12 believed that they worked best with the thumbnail layout.

4.3 Lymphatic Vessel Identification Task

While we only discovered the contextual use of information in the Size estimation task during the experiment, this task was designed to encourage use of information across images. For each set, the subjects drew the walls of the lymphatic vessel. Since we are interested in the boundaries of the wall, this approach yielded two responses for each set. We treated the two walls of the vessel independently and accepted answers where the right wall of the vessel was identified as the left wall, as the subject had believed that the vessel was parallel to its actual location.

As illustrated in figure 4, the experts performed better than the novices in correctly identifying walls of the lymphatic vessels. While the experts identified 75.3% of the vessel boundaries across all the layouts, the novices could only identify 51% (p=0.03). When comparing success on this task by the type of view, the experts' performance across the layouts varied significantly (p=0.035). They returned the best results for Montage (87%) and the worst for Thumbnail (63%). This was somewhat expected as the montage layout acted as a slowed down video interface and facilitated the comparison of differences in a set of somewhat similar images. However, the novices surprised us by scoring the lowest with this layout (44%) and the highest with the thumbnail view (55%) which we expected to be the most ill suited layout for this task. This result contradicted our expectation that the novices and experts alike would benefit the most from the video-like montage layout. In spite of their performance, 4 of the 6 novices (11 of 15 subjects overall) preferred the montage layout and 3 of these believed that they worked better with this layout.

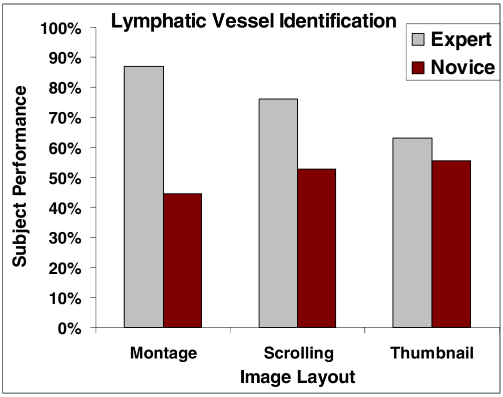


Fig. 4. Performance of experts and novices on the Lymphatic vessel identification task

5 Discussion

Familiarity with the image layouts may have influenced the subjects' choices for the most-liked and most-effective interfaces. While the subjects had not worked with the scrolling layout before, they adapted to this view quickly and their performance on this view was not significantly worse than the other layouts.

Time management also played a role in user performance with the different layouts. The thumbnail layout allowed users to manage their time freely and devote it to those images that they found challenging. However, this did not turn out to be the most effective test-taking strategy as some subjects lost track of the available time and consequently had to forego answering some questions. In contrast, the other views naturally constrained the time available per image and aided the subjects in managing their time.

Furthermore, while most interfaces for personal photograph presentations employ external contextual information, such as the timestamp [14] or the geographic location [10] the contextual information for the tasks of our interest is derived directly from other images that are displayed simultaneously. Our tasks required subjects to use this contextual information in two ways: in the Size estimation task, subjects used a few images as a baseline for estimating the size of objects in other images, much like unordered reference points; for the Lymphatic Vessel Wall identification, they used the images in sequence and focused on the differences between sequential images. Our tasks also require specialized training and the subjects are typically more interested in the characteristics of images and their content, unlike personal collections where users typically search for familiar images.

Our strategy of using still images to test for video-acquired skills seems somewhat analogous to the extraction of key frames from video [23]. The key frames are expected to convey a sense of the larger collection, in this case, the video from where they are extracted. This is somewhat similar to the concept of collection understanding [7]. Christel, et al., have explored the use of various video abstractions for key frame representation [8]. However, our images do not represent (sections of) videos; we are interested in the cognitive decisions of our users based upon the details within these images. Our technique has more in common with the researchers who slowed down videos of honeybees to analyze the still frames and discover the nature of their flight [2].

Our research contributes to the areas of Digital Libraries and Medical Imaging Systems. Prior research at the intersection of these areas has focused on digital collections for augmenting the virtual laboratory [3] and support for querying over medical information [16]. Other research for imparting scientific training has explored graphics intensive techniques to generate high-quality virtual reality training simulations [9, 22]. Artificial Intelligence-enhanced tutoring environments have also been designed for training professionals in medical areas [21].

6 Future Work

Currently we are exploring mechanisms to shorten the training period for new researchers in order to make their semester-long research experience richer and more productive. This is critical since each semester the bat lab trains a large number of new undergraduate and graduate students in cardiovascular research methods. Typically, new students work with experienced researchers and acquire the necessary skills over the course of the semester. Lab equipment is a critical resource and is only sparingly available to trainees. The semester-long training presents a significant barrier when attracting young researchers and only a few continue their research into the following semesters. An understanding of the interfaces that novice and experienced researchers can work with gained from our study will aid us in designing these training systems for developing novice researchers into experts. Furthermore, since the three layouts that we selected for the purpose of this study are by no means the only candidates for information displays for cognitive tasks. We are currently exploring the contextual properties of other layouts.

In addition, while our subjects can intuit the nature of the blood vessels that they are dealing with, this did not translate into ability to accurately estimate the diameter of these vessels. Clearly, researchers can benefit from additional cues to aid them in accurately applying known real world dimensions to scaled photographs. The nature of the additional information that is necessary for such translations and the mechanisms for delivering this information unambiguously, yet unobtrusively, are significant challenges that merit investigation.

Acknowledgements. Jay Widmer, Arun Venugopal, and Missy Young were instrumental in the development of the image corpus.

References

1. AltaVista - Image Search. 2006. http://images.altavista.com/ viewed Feb. 2006.
2. Altshuler, D., Dickson, W., Vance, J., Roberts, S. and Dickinson, M. 2005. Short-amplitude high-frequency wing strokes determine the aerodynamics of honeybee flight. PNAS, Dec 2005; 102: pp. 18213 - 18218.
3. Bartolo, L., Lowe, C., Sadoway, D., and Trapa, P. 2005. Large Introductory Science Courses & Digital Libraries. In Proc. of JCDL 2005 (Denver, CO) ACM Press, pp. 366.
4. Bederson, B. 2001. PhotoMesa: A Zoomable Image Browser Using Quantum Treemaps and Bubblemaps. In Proc. of UIST ’01 (Orlando, FL) ACM Press, pp. 71-80.
5. Börner, K., Dillon, A., Dolinsky, M. 2000. LVis—Digital Library Visualizer. In Proc. of IEEE InfoViz 2000 (London UK) IEEE Press, pp. 77-81.
6. Corbis: stock photography and digital pictures. 2006. http://pro.corbis.com/ viewed Feb. 2006.
7. Chang, M., Leggett, J., Furuta, R., Kerne, A., Williams, J., Burns, S., and Bias, R. 2004. Collection understanding. In Proc. of JCDL 2004 (Tuscon, AZ) ACM Press, pp. 334-342.
8. Christel, M., Winkler, D., and Taylor, C. 1997. Multimedia abstractions for a digital video library. In Proc. of DL ’97 (Philadelphia, PA). ACM Press, pp. 21-29.
9. de Lima, L., Nunes, F., Takashi, R., Rodello, I., Brega, J., and Sementille, A. 2004. Virtual Reality for Medical Training: A Prototype to Simulate Breast Aspiration Exam. In Proc. of VRCAI ’04 (Singapore) ACM Press, pp. 328-331.
10. Frappr! – Group Maps for Online Communities. 2006. http://www.frappr.com/ viewed Mar. 2006.
11. Graham, A., Garcia-Molina, H., Paepcke, A., and Winograd, T. 2002. Time as Essence for Photo Browsing Through Personal Digital Libraries. In Proc. of JCDL 2002 (Portland, OR) ACM Press, pp. 326-335.
12. Google Image Search. 2006. http://images.google.com/ viewed February 2006.
13. Greenberg, S. and Rounding, M. 2001. The Notification Collage: Posting Information to Public and Personal Displays. In Proc. Of CHI ’01 (Seattle, WA) ACM Press, pp. 514-521.

14. Harada, S., Naaman, M., Song, Y., Wang, Q., and Paepcke, A. 2004. Lost in Memories: Interacting with Photo Collections on PDAs. In Proc. of JCDL 2004 (Tuscon, AZ) ACM Press, pp. 325-333.
15. Kerne, A. 2001. CollageMachine: Interest-Driven Browsing Through Streaming Collage. Proc Cast01: Living in Mixed Realities. Bonn, Germany, pp. 241-244.
16. Kholief, M., Maly, K., and Shen, S. 2003. Event-based Retrieval from a Digital Library Containing Medical Streams. In Proc. of JCDL 2003 (Houston, TX) ACM Press, pp. 231-233.
17. Kuchinsky, A., Pering, C., Creech, M., Freeze, D., Serra, B., and Gwizdka, J. 1999. FotoFile: A Consumer Multimedia Organization and Retrieval System. In Proceedings of CHI ’99 (Pittsburgh PA), ACM Press, pp. 496-503.
18. Kustanowitz, J. and Shneiderman, B. 2005. Meaningful Presentations of Photo Libraries: Rationale and Applications of Bi-level Radial Quantum Layouts. In Proc. of JCDL 2005 (Denver, CO) ACM Press, pp. 188-196.
19. NYPL Digital Gallery. 2006. http://digitalgallery.nypl.org/nypldigital/ viewed Feb. 2006.
20. Rodden, K., and Wood, K 2003. How Do People Manage Their Digital Photographs? In Proc. of CHI 2003 (Ft. Lauderdale, FL) ACM Press, pp. 409-416.
21. Suebnukarn, S. and Haddawy, P. 2004. A Collaborative Intelligent Tutoring System for Medical Problem-based Learning. In Proc. of IUI ’04 (Funchal, Madeira) ACM Press, pp. 14-21.
22. Wang, Z., Chui, C., Cai, Y., and Ang, C. 2004. Multidimensional Volume Visualization for PC-based Microsurgical Simulation System. In Proc. of VRCAI '04 (Singapore) ACM Press, pp. 309-316.
23. Wildemuth, B. M., Marchionini, G., Yang, M., Geisler, G., Wilkens, T., Hughes, A., and Gruss, R. 2003. How Fast is Too Fast?: Evaluating Fast Forward Surrogates for Digital Video. In Proc. of JCDL 2003 (Houston, TX) ACM Press, pp. 221-230.
24. Yahoo! Photos. 2006. http://photos.yahoo.com/ viewed Feb. 2006.

Incorporating Cross-Document Relationships Between Sentences for Single Document Summarizations

Xiaojun Wan, Jianwu Yang, and Jianguo Xiao

Institute of Computer Science and Technology,
Peking University, Beijing 100871, China
{wanxiaojun, yangjianwu, xiaojianguo}@icst.pku.edu.cn

Abstract. Graph-based ranking algorithms have recently been proposed for single document summarizations and such algorithms evaluate the importance of a sentence by making use of the relationships between sentences in the document in a recursive way. In this paper, we investigate using other related or relevant documents to improve summarization of one single document based on the graph-based ranking algorithm. In addition to the within-document relationships between sentences in the specified document, the cross-document relationships between sentences in different documents are also taken into account in the proposed approach. We evaluate the performance of the proposed approach on DUC 2002 data with the ROUGE metric and results demonstrate that the cross-document relationships between sentences in different but related documents can significantly improve the performance of single document summarization.

1 Introduction

Text summarization is the process of automatically creating a compressed version of a given text that provides useful information for users. Automated text summarization has drawn much attention in recent years because it becomes more and more important in many text applications. For example, current search engines usually provide a short summary for each resultant document so as to facilitate users to browse the results and improve users' search experience. News agents usually provide concise headline news describing hot news and they also produce weekly news review for users, which saves users' time and provide better service quality.

Text summaries can be either query-relevant summaries or generic summaries. A query-relevant summary is usually used in search engines and its content should be closely related to the given query. And a generic summary should contain the main topics of the document while keeping redundancy to a minimum. It is a great challenge to automatically generate a high-quality generic summary for a document without any additional clues and prior knowledge. In this paper, we focus on generic single document summarization.

To the best of our knowledge, almost all previous methods for single document summarization produce a summary for a specified document based only on the information contained in the document. In some cases, a set of related or relevant documents are provided and some single documents in the set are required to be

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 403 – 414, 2006.

summarized. For example, the documents returned by a search engine for a specified query can be considered topically related to each other. The documents within a cluster produced by a clustering algorithm on a document set are also deemed related and relevant. This study aims to explore whether the cross-document relationships between sentences in different but related documents can contribute to the task of single document summarization. In this paper, we propose the novel idea of incorporating both the cross-document relationships between sentences and the within-document relationships between sentences into the graph-based ranking algorithm for single document summarization. By taking into account these two kinds of relationships between sentences, each sentence in a single document obtains a global ranking score to denote its information richness. Then a greedy algorithm is employed to impose diversity penalty on each sentence of the document based on the overlap between this sentence and other high informative sentences in the document. The sentences with both high information richness and high information novelty are chosen into the single summary for the specified document. We perform experiments on DUC 2002 data and experimental results show that the cross-document relationships between sentences can significantly improve the performance of single document summarization.

The rest of this paper is organized as follows: Section 2 briefly introduces related work. The details of the proposed approach are described in Section 3. Section 4 presents and discusses the evaluation results. Lastly we conclude our paper in Section 5.

2 Related Work

In recent years, single document summarization has been widely explored in the natural language processing and information retrieval communities. A series of workshops and conferences on automatic text summarization (e.g. SUMMAC[1], DUC[2] and NTCIR[3]), special topic sessions in ACL, COLING, and SIGIR have advanced the technology and produced a couple of experimental online systems.

Generally speaking, single document summarization methods can be categorized into two categories: extraction-based methods and abstraction-based methods [9, 10, 13]. Extraction is much easier than abstraction because extraction is just to select existing sentences while abstraction needs sentence compression and reformulation. In this paper, we focus on extraction-based methods.

Extraction-based methods usually assign each sentence a saliency score and then rank the sentences in the document. The scores is usually assigned based on a combination of statistical and linguistic features, including term frequency [17], sentence position [8], cue words [5], stigma words [5], topic signature [16], lexical chains [22], etc. Machine learning methods are also employed to extract sentences, including classification-based methods [1, 14], clustering-based methods [21], HMM-based methods [4], etc. Other methods derived from information retrieval techniques is developed for sentence extraction, including maximal marginal relevance (MMR) [3], latent semantic analysis (LSA) [7], and relevance measure [7].

[1] http://www.itl.nist.gov/iaui/894.02/related_projects/tipster_summac/

[2] http://duc.nist.gov

[3] http://research.nii.ac.jp/ntcir/index-en.html

In [23], the mutual reinforcement principle is employed to iteratively extract key phrases and sentences from a document. Moreover, a method based on text segmentation is proposed by McDonald and Chen [18] and the text segments instead of the sentences are ranked.

Most recently, graph-based ranking methods, including TextRank [19, 20] and LexPageRank [6] have been proposed for document summarization. Similar to PageRank [2] or HITS [12], these methods first build a graph based on the similarity relationships between sentences in a document and then the importance of a sentence is determined by taking into account global information on the graph recursively, rather than relying only on local sentence-specific information. The basic idea underlying the graph-based ranking algorithm is that of "voting" or "recommendation". When one sentence links to another one, it is basically casting a vote for that other sentence. The higher the number of votes that are cast for a sentence, the higher the importance of the sentence. Moreover, the importance of the sentence casting the vote determines how important the vote itself is. The computation of sentence importance is usually based on a recursive form, which can be transformed into the problem of solving the principal eigenvector of the transition matrix.

While in the above graph-based ranking algorithms, each single document is summarized independently, in other words, only sentences within the same document cast votes for each other. We believe that the sentences in other related documents can also cast votes for the sentences in the specified document because for a set of related documents, the information contained in an important sentence of a document will be expressed in other sentences of the other documents. Moreover, if needed, our approach can summarize all single documents in the document set in a batch way.

3 The Proposed Approach

The proposed approach summarizes each single document within a document set based on the graph-based ranking algorithm over all sentences in the document set. The documents in the document set are assumed to be related or relevant[4]. The contribution of the proposed approach is based on the following intuition: The important information expressed in a sentence of a document is also expressed in the sentences of many related documents besides the other sentences within the same document. Figure 1 gives the framework of the proposed approach.

In the framework, the first step aims to build a global affinity graph reflecting the relationships among all sentences in the document set; the second step is to compute information richness of each sentence based on the global affinity graph. The first two steps performs on the whole document set, while the third step performs only on the single document, in other words, the process of information richness computation is on a document set scale and the process of diversity penalty is on a single document scale. We assume that a good summary is expected to include the sentences with both high information richness and high information novelty.

[4] As noted in Section 1, we can obtain a set of related documents by clustering algorithms or information retrieval techniques.

```
1. Build a global affinity graph G based on all sentences in the document set
   $D=\{d_1,d_2,\ldots d_l\}$. Let the $S=\{s_1, s_2, \ldots, s_n\}$ denotes the sentence set.
2. Based on the global affinity graph G, the graph-based ranking algorithm is
   employed to compute a global ranking score $InfoRich(s_i)$ for each sentence
   $s_i$, where $InfoRich(s_i)$ denotes the information richness of the sentence $s_i$.
3. for any single document $d_k$ to be summarized
      1) Extract the local affinity graph $G_{d_k}$ for $d_k$ from G; Let $S_{d_k}$ denotes
      the set of sentences in $d_k$.
      2) Impose a diversity penalty on each sentence in $S_{d_k}$ based on $G_{d_k}$ and
      the obtained global ranking scores of sentences in $S_{d_k}$, and obtain a
      overall affinity ranking score $ARScore(s_i)$ for each sentence $s_i$ in $S_{d_k}$ .
      3)Choose the sentences with highest overall ranking scores into the
      summary;
```

Fig. 1. The framework of the proposed approach

3.1 Global Affinity Graph Building

Given the sentence collection $S=\{s_i \mid 1 \leqslant i \leqslant n\}$, we measure the similarity between sentences based on co-occurrences of terms in the sentences. Formally, a sentence s_i is represented by the set of N_i words that appear in the sentence: $s_i = w_1^i, w_2^i, \ldots, w_{N_i}^i$. Given two sentences s_i and s_j, the similarity of s_i and s_j is defined as:

$$\mathrm{sim}(s_i, s_j) = \frac{\left|\{w_k \mid w_k \in s_i \ \& \ w_k \in s_j\}\right|}{\left|s_i\right| + \left|s_j\right| - \left|\{w_k \mid w_k \in s_i \ \& \ w_k \in s_j\}\right|} \tag{1}$$

The above measure is known as the Jaccard coefficient [11]. Other sentence similarity measures, such as Cosine similarity, Overlap coefficient, Dice coefficient, etc. are also possible, and we are currently evaluating their impact on the summarization performance.

If sentences are considered as nodes, the sentence collection can be modeled as an undirected graph by generating the edge (link) between two sentences only if their similarity weight exceeds 0, i.e. an undirected link between s_i and s_j ($i \neq j$) with similarity weight $sim(s_i, s_j)$ is constructed if $sim(s_i,s_j)>0$; otherwise no link is constructed.

Thus, we construct an undirected graph G reflecting the relationships between sentences by their content similarity. The graph is called as Affinity Graph. Since the graph contains all sentences in the document set, it is called as Global Affinity Graph.

3.2 Information Richness Computation

The graph-based ranking algorithm [6, 19, 20] is employed to compute information richness of sentences, which is based on the following three intuitions:

1. The more neighbors a sentence has, the more informative it is;
2. The more informative a sentence's neighbors are, the more informative it is.
3. The more heavily a sentence is linked with other informative sentences, the more informative it is.

In previous graph-based ranking algorithms for single document summarization, the neighbors of a sentence all come from the same document, while it is intuitive that the information contained in an informative sentence will be also expressed in the sentences of other related documents and we believe that the votes of neighbors in related documents are also important, so we use both the neighbors from the same document and the neighbors from related documents to iteratively compute the information richness of a sentence.

The graph-based ranking algorithm is similar to PageRank [2]. First, we use an adjacency (affinity) matrix **M** to describe the affinity graph with each entry corresponding to the weight of a link in the graph. **M** = $(M_{i,j})_{n\times n}$ is defined as follows:

$$M_{i,j}=\begin{cases}\mathrm{sim}(s_i,s_j), & \text{if } i\neq j\\ 0 & \text{, otherwise}\end{cases} \tag{2}$$

In our context, the links (edges) between sentences in the graph can be categorized into two classes: intra-document link and inter-document link. Given a link between a sentence pair of s_i and s_j, if s_i and s_j come from the same document, the link is called an intra-document link; and if s_i and s_j come from different documents, the link is called an inter-document link. We believe that intra-document links and inter-document links have unequal contributions in the graph based ranking algorithm, so distinct weights are assigned to intra-document links and inter-document links respectively. We decompose the original affinity matrix **M** as

$$\mathbf{M}=\mathbf{M}_{\text{intra}}+\mathbf{M}_{\text{inter}} \tag{3}$$

where $\mathbf{M}_{\text{intra}}$ is the affinity matrix containing only the intra-document links (the entries of inter-document links are set to 0) and $\mathbf{M}_{\text{inter}}$ is the affinity matrix containing only the inter-document links (the entries of intra-document links are set to 0).

After we differentiate the intra-document links and inter-document links, the new affinity matrix is as follows:

$$\widehat{\mathbf{M}}=\lambda_1\mathbf{M}_{\text{intra}}+\lambda_2\mathbf{M}_{\text{inter}} \tag{4}$$

We let $\lambda_1,\lambda_2\in[0,1]$ in the experiments. If λ_1 =0 and λ_2 =1, only inter-document links are taken into account in the algorithm, and if λ_1 =1 and λ_2 =0, only intra-document links are taken into account in the algorithm. Note that if $\lambda_1=\lambda_2$ =1, Equation (4) reduces to Equation (3).

Then $\overset{\frown}{\mathbf{M}}$ is normalized as follows to make the sum of each row equal to 1:

$$\tilde{M}_{i,j} = \begin{cases} \overset{\frown}{M}_{i,j} \Big/ \sum_{j=1}^{n} \overset{\frown}{M}_{i,j}, & \text{if } \sum_{j=1}^{n} \overset{\frown}{M}_{i,j} \neq 0 \\ 0 & \text{, otherwise} \end{cases} \tag{5}$$

Note that now we do not have $\tilde{M}_{i,j} = \tilde{M}_{j,i}$ for any pair of *i* and *j*. Based on the normalized adjacency matrix $\tilde{\mathbf{M}} = (\tilde{M}_{i,j})_{n\times n}$, the information richness score for each node can be deduced from those of all other nodes linked with it and it can be formulated in a recursive form as follows:

$$\text{InfoRich}(s_i) = \sum_{\text{all } j\neq i} \text{InfoRich}(s_j) \cdot \tilde{M}_{j,i} \tag{6}$$

The above form can be represented in a matrix form:

$$\vec{\lambda} = \tilde{\mathbf{M}}^{T} \vec{\lambda} \tag{7}$$

where $\vec{\lambda} = [\text{InfoRich}(s_i)]_{n\times 1}$ is the eigenvector of $\tilde{\mathbf{M}}^{T}$.

Note that $\tilde{\mathbf{M}}$ is normally a sparse matrix and some rows with all-zero elements could possibly appear because some sentences have no links with other sentences. Similar to the random jumping factor in PageRank, a damping factor *d* (usually 0.85) is introduced in order to compute a reasonable eigenvector:

$$\text{InfoRich}(s_i) = d \cdot \sum_{\text{all } j\neq i} \text{InfoRich}(s_j) \cdot \tilde{M}_{j,i} + \frac{(1-d)}{n} \tag{8}$$

And the matrix form is:

$$\vec{\lambda} = d\tilde{\mathbf{M}}^{T} \vec{\lambda} + \frac{(1-d)}{n} \vec{e} \tag{9}$$

where $\vec{e}$ is a unit vector with all elements equaling to 1.

The above process can be considered as a Markov chain by taking the sentences as the states and the corresponding transition matrix is given by $d\tilde{\mathbf{M}}^{T} + (1-d)\mathbf{U}$, where $\mathbf{U} = [\frac{1}{n}]_{n\times n}$. The stationary probability distribution of each state is obtained by the principal eigenvector of the transition matrix.

3.3 Diversity Penalty Imposition

After the information richness of each sentence is computed based on the global affinity graph, we can choose highly informative sentences into the summary for any specified single document in the document set. However, a good summary should keep redundant information as minimal as possible, so we impose a diversity penalty to each sentence. Finally, an overall affinity rank score is obtained to reflect both information richness and information novelty of a sentence in the specified document. Since we aim to produce single document summaries, this diversity penalty process must be applied for each single document separately.

For each single document d_k to be summarized we can extract a sub-graph G_{d_k} only containing the sentences within d_k and the corresponding edges between them from the global affinity graph G. We assume the document d_k has m ($m<n$) sentences and the sentences' affinity matrix $\mathbf{M}_{\mathrm{d_k}} = (\mathbf{M}_{\mathrm{d_k}})_{\mathrm{m \times m}}$ is derived from the original matrix $\mathbf{M}$ by extracting the corresponding entries. Then $\mathbf{M}_{\mathrm{d_k}}$ is normalized into $\tilde{\mathbf{M}}_{\mathrm{d_k}}$ as Equation (5) to make the sum of each row equal to 1. Similar to [24], a greedy algorithm is used to impose the diversity penalty and compute the final affinity rank score for each sentence within the document. The algorithm goes as follows:

1. *Initialize two sets A=*ϕ*, B={*s_i *| i=1,2,...,m} for the specified document* d_k*, and each sentence's overall affinity rank score is initialized to its information richness score, i.e. ARScore(*s_i*) = InfoRich(*s_i*), i=1,2,...m.*
2. *Sort the sentences in B by their current affinity rank scores in descending order.*
3. *Suppose* s_i *is the highest ranked sentence, i.e. the first sentence in the ranked list. Move sentence* s_i *from B to A, and then a diversity penalty is imposed to the affinity rank score of each sentence linked with* s_i *in B as follows:*

   For each sentence $s_j \in B$

   $ARScore(s_j) = ARScore(s_j) - \omega \cdot (\tilde{M}_{\mathrm{d_k}})_{j,i} \cdot InfoRich(s_i)$

   where ω>0 *is the penalty degree factor. The larger* ω *is, the greater penalty is imposed to the affinity rank score. If* ω*=0, no diversity penalty is imposed at all.*
4. *Go to step 2 and iterate until B=*ϕ *or the iteration count reaches a predefined maximum number.*

Fig. 2. The algorithm of diversity penalty imposition

In the above algorithm, the third step is the crucial step and its basic idea is to decrease the affinity rank score of less informative sentences by the part conveyed from the most informative one. After the affinity rank scores are obtained for all m sentences in the document d_k, several sentences with highest affinity rank scores are chosen to produce the summary for d_k according to the summary length limit.

The above algorithm is applied once for each single document to be summarized in the document set.

4 Experiments

4.1 Data Set

Single document summarization has been one of the fundamental tasks in DUC 2001 and DUC 2002, i.e. task 1 of DUC 2001 and task 1 of DUC 2002. We used DUC 2001 data for training and DUC 2002 data for testing in the experiments. The task 1 of DUC 2002 aims to evaluate generic summaries with a length of approximately 100 words or less. DUC 2002 provides 567 English news articles for single-document summarization task. The sentences in each article have been separated and the sentence information is stored into files. The 567 articles are collected from TREC-9 and grouped into 59 clusters[5] and the documents within each cluster are related or relevant, so it is appropriate to apply our proposed approach directly. The single summaries for all documents within a cluster are produced in a batch way.

As a preprocessing step, for each document, the dialog sentences (sentences in quotation marks) were removed. The stop words in each sentence were removed and the remaining words were stemmed using the Porter's stemmer[6].

4.2 Evaluation Metric

We use the ROUGE [15] evaluation toolkit[7] for evaluation, which is adopted by DUC for automatic summarization evaluation. It measures summary quality by counting overlapping units such as the n-gram, word sequences and word pairs between the candidate summary and the reference summary. ROUGE-N is an n-gram recall measure computed as follows:

$$\mathrm{ROUGE-N} = \frac{\sum_{S\in\{\mathrm{Ref\,Sum}\}} \sum_{\mathrm{n\text{-}gram}\in S} \mathrm{Count}_{\mathrm{match}}(\mathrm{n-gram})}{\sum_{S\in\{\mathrm{Ref\,Sum}\}} \sum_{\mathrm{n\text{-}gram}\in S} \mathrm{Count}(\mathrm{n-gram})} \tag{10}$$

where *n* stands for the length of the n-gram, and *Count*$_{match}$*(n-gram)* is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries. *Count(n-gram)* is the number of n-grams in the reference summaries.

ROUGE toolkit reports separate scores for 1, 2, 3 and 4-gram, and also for longest common subsequence co-occurrences. Among these different scores, unigram-based ROUGE score (ROUGE-1) has been shown to agree with human judgment most [15]. We show three of the ROUGE metrics in the experimental results, at a confidence level of 95%: ROUGE-1 (unigram-based), ROUGE-2 (bigram-based), and ROUGE-W (based on weighted longest common subsequence, weight=1.2). Note that we mainly concern ourselves with ROUGE-1 scores.

In order to truncate summaries longer than 100 words, we use the "-l 100" option[8] in ROUGE toolkit and we also use the "-m" option for word stemming.

[5] At first, there were 60 document clusters, but the document cluster of D088 is withdrawn by NIST due to differences in documents used by systems and NIST summarizers.

[6] http://www.tartarus.org/martin/PorterStemmer/

[7] We use ROUGEeval-1.4.2 downloaded from http://haydn.isi.edu/ROUGE/

[8] This option is necessary for fair comparison because longer summary will usually increase ROUGE evaluation scores.

4.3 Experimental Results

4.3.1 System Comparison

In the experiments, the proposed system has been compared with top 5 (out of 15) systems and baseline systems. The top five systems are the systems with highest ROUGE scores, chosen from the performing systems on the single document summarization task of DUC 2002. Table 1 shows the system comparison results over three ROUGE metrics[9]. In the table, S21-S31 are the system IDs for the top performing systems. "Intra- & Inter-document link" denotes the proposed approach taking into account both intra-document links between sentences within the specified document and inter-document links between sentences across different but related documents. "Only Inter-document link" and "Only Intra-document link" are two baseline systems: the first one is based only on inter-document links and the second one is based only on intra-document links. Note that previous summarization work [6, 19, 20] using graph-based ranking algorithm is similar to "Only Intra-document link" in this paper. The performance of "Intra- & Inter-document link" is achieved when $\lambda_1=1$ and $\lambda_2=0.7$, $\omega=1$. The performance of "Only Inter- document link" is achieved when $\lambda_1=0$ and $\lambda_2=1$, $\omega=1$. And the performance of "Only Intra-document link" is achieved when $\lambda_1=1$ and $\lambda_2=0$, $\omega=0.5$. Note that the parameters are tuned on DUC 2001 data.

Table 1. System comparison on DUC 2002 data

System	ROUGE-1	ROUGE-2	ROUGE-W
S28	0.48049	0.22832	0.17073
S21	0.47754	0.22273	0.16814
Intra- & Inter-document link	0.47710	0.20457	0.16344
Only Inter-document link	0.47399	0.20332	0.16215
S31	0.46506	0.20392	0.16162
Only Intra-document link	0.46443	0.19072	0.15832
S29	0.46384	0.21246	0.16462
S27	0.46019	0.21273	0.16342

Seen from the table, our proposed system, i.e. "Intra- & Inter-document link", achieves a good performance comparable to that of the state-of-the-art systems, i.e. S28 and S21. The proposed system outperforms both the system based only on the intra-document links (i.e. "Only Intra-document link") and the system based only on the inter-document links (i.e. "Only Inter-document link"), which demonstrates that both the intra-document links and the inter-document links between sentences are important for single document summarization based on the graph-based ranking algorithm. We can also see that the system based only on the inter-document links (i.e. "Only Inter-document link") outperforms the system based only on the intra-

[9] The ROUGE values of top performing systems are different from those reported in [19, 20] because they do not use the "-l 100" option to truncate summaries longer than 100 words.

document links (i.e. "Only Intra-document link"), which further demonstrates the great importance of the cross-document relationships between sentences for single document summarization.

4.3.2 Parameter Tuning

In this section, we investigate tuning the important parameters employed in the proposed systems, including the penalty factor ω for three systems based on graph ranking algorithms, the intra-document link and inter-document link differentiation weights λ_1 and λ_2 for the proposed system, i.e. "Intra- & Inter-document link". The ROUGE-1 results are shown in Figures 3-4 respectively.

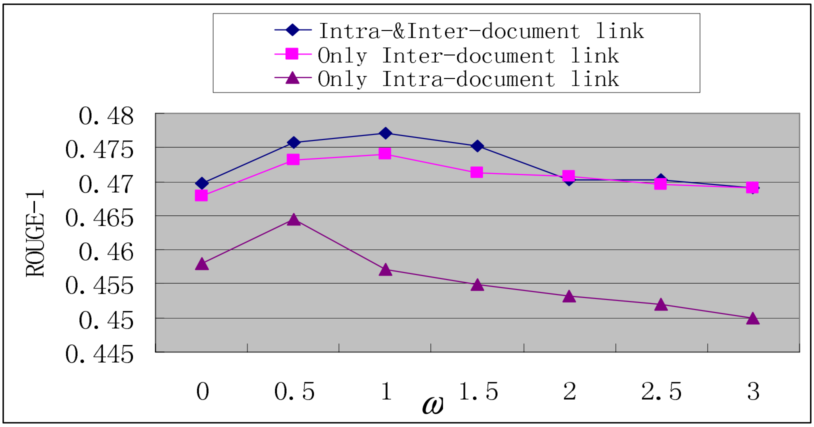


Fig. 3. Penalty factor (ω) tuning for three systems

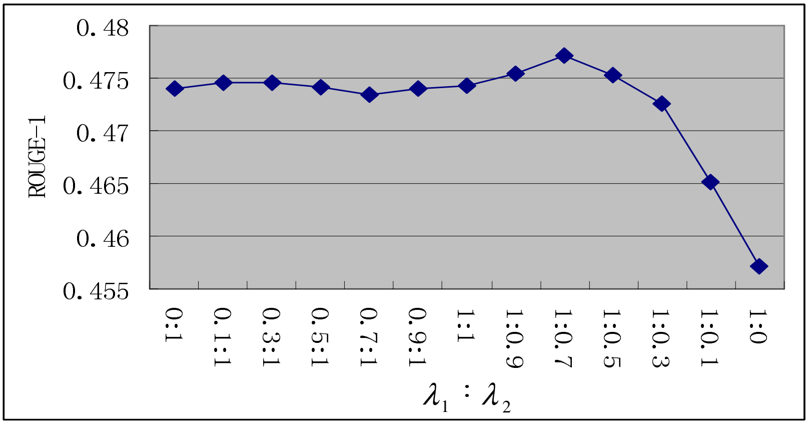


Fig. 4. Intra-document/inter-document link weight ($\lambda_1 : \lambda_2$) tuning for the proposed system (i.e. "Intra- & Inter-document link")

Figure 3 demonstrates the influence of the penalty factor ω for the proposed system when λ_1=1 and λ_2=0.7, and also for the systems of “Only Intra-document link” and “Only Inter-document link”. It shows that the proposed system outperforms the two baseline systems over different values of the penalty factor ω. Moreover, the system of “Only Inter-document link” much outperforms the system of “Only Intra-document link” irrespective of the value of ω. We can also see that ω=1 is the point where the proposed system and the system of “Only Inter-document link” achieve their best performances, and more or less diversity penalty will deteriorate their performances.

Figure 4 shows the influence of the intra-document/inter-document link weights λ_1 and λ_2 for the proposed system when ω=1. λ_1 and λ_2 range from 0 to 1. In Figure 4, $\lambda_1 : \lambda_2$ denotes the real values λ_1 and λ_2 are set to. For example, $\lambda_1 : \lambda_2$=1:1 means λ_1=1 and λ_2=1. It is observed that when λ_1=0.3 and λ_2=1 the system can obtain the optimal performance.

5 Conclusion

In this paper, we propose to incorporate cross-document relationships between sentences into the graph-based ranking algorithm for single document summarizations. Experimental results on DUC 2002 data demonstrate the great importance of inter-document links between sentences in different but related documents for single document summarizations based on graph-based ranking algorithm.

References

1. Amini, M. R., Gallinari, P.: The Use of Unlabeled Data to Improve Supervised Learning for Text Summarization. In Proceedings of SIGIR2002, 105-112.
2. Brin, S. and Page, L. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 1998, 30:1-7.
3. Carbonell, J., Goldstein, J.: The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998, 335-336.
4. Conroy, J. M., O’Leary, D. P.: Text Summarization via Hidden Markov Models. In Proceedings of SIGIR2001, 406-407.
5. Edmundson, H. P.: New Methods in Automatic Abstracting. Journal of the Association for computing Machinery, 1969, 16(2): 264-285.
6. ErKan, G¨unes, Radev, D. R.: LexPageRank: Prestige in Multi-Document Text Summarization. In Proceedings of EMNLP2004.
7. Gong, Y. H., Liu, X.: Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. In Proceedings of SIGIR2001, 19-25.
8. Hovy, E., Lin, C. Y.: Automated Text Summarization in SUMMARIST. In Proceeding of ACL’1997/EACL’1997 Worshop on Intelligent Scalable Text Summarization, 1997.

9. Jing, H.: Sentence Reduction for Automatic Text Summarization. In Proceedings of ANLP 2000.
10. Jing, H., McKeown, K. R.: Cut and Paste Based Text Summarization. In Proceedings of NAACL2000, 178-185.
11. Jones, W. P. and Furnas, G. W. Pictures of relevance: a geometric analysis of similarity measure. Journal of the American Society for Information Science, 1987, 38(6): 420-442.
12. Kleinberg, J. M. Authoritative sources in a hyperlinked environment. Journal of the ACM, 1999, 46(5):604-632.
13. Knight, K., Marcu, D.: Summarization beyond Sentence Extraction: A Probabilistic Approach to Sentence Compression. Artificial Intelligence, 2002, 139(1): 91-107.
14. Kupiec, J., Pedersen, J., Chen, F.: A.Trainable Document Summarizer. In Proceedings of SIGIR1995, 68-73.
15. Lin, C. Y., Hovy, E.: Automatic Evaluation of Summaries Using N-Gram Co-Occurrence Statistics. In Proceedings of HLT-NAACL2003.
16. Lin, C. Y., Hovy, E.: The Automated Acquisition of Topic Signatures for Text Summarization. In Proceedings of the 17th Conference on Computational Linguistics, 2000, 495-501.
17. Luhn, H. P.: The Automatic Creation of literature Abstracts. IBM Journal of Research and Development, 1969, 2(2).
18. McDonald, D., Chen, H.: Using Sentence-Selection Heuristics to Rank Text Segment in TXTRACTOR. In Proceedings of JCDL2002, 28-35.
19. Mihalcea, R., Tarau, P.: TextRank: Bringing Order into Texts. In Proceedings of EMNLP2004.
20. Mihalcea, R. and Tarau, P.: A language independent algorithm for single and multiple document summarization. In Proceedings of IJCNLP2005.
21. Nomoto, T., Matsumoto, Y.: A New Approach to Unsupervised Text Summarization. In Proceedings of SIGIR2001, 26-34.
22. Silber, H. G., McCoy, K.: Efficient Text Summarization Using Lexical Chains. In Proceedings of the 5th International Conference on Intelligent User Interfaces, 2000, 252-255.
23. Zha, H. Y.: Generic Summarization and Keyphrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering. In Proceedings of SIGIR2002, 113-120.
24. Zhang, B., Li, H., Liu, Y., Ji, L., Xi, W., Fan, W., Chen, Z., and Ma, W.-Y. Improving web search results using affinity graph. In Proceedings of SIGIR2005.

Effective Content Tracking for Digital Rights Management in Digital Libraries

Jen-Hao Hsiao, Cheng-Hung Li, Chih-Yi Chiu, Jenq-Haur Wang,
Chu-Song Chen, and Lee-Feng Chien

Institute of Information Science, Academia Sinica, Taipei, Taiwan
{jenhao, chli, cychiu, jhwang, song, lfchien}@iis.sinica.edu.tw

Abstract. A usual way for content protection of digital libraries is to use digital watermarks and a DRM-based access-control environment. These methods, however, have limitations. Digital watermarks embedded in digital content could be removed by malicious users via post-processing, whereas DRM-based access-control solutions could be hacked. In this paper, we introduce a content tracking mechanism that we have built for multimedia-content near-replica detection as the second line of defense. The integrated framework aims to detect unlawful copyright infringements on the Internet, and combines the strengths of static rights enforcement and dynamic illegal content tracking. The issues of accuracy and huge computation cost in copy detection have been addressed by the introduced content-based techniques. Our experiments demonstrate the efficacy of proposed copy detector.

1 Introduction

Protection of the copyrights and revenues of content owners in digital libraries has become increasingly important in recent years. Since digital content differs from objects in real world, it can be easily copied, altered, and distributed to a large number of recipients. This almost certainly causes copyright infringement and therefore revenue losses to content owners. The National Digital Archives Program (NDAP) of Taiwan has amassed a rich collection of cultural and historical artifacts. These assets have been digitized to enhance their preservation, and make them more accessible to users. The metadata and digital content storage systems are called archival systems, and – like other types of digital content – they too face the problem of piracy. Thus, content holders are sometimes unwilling to release digital content, because their intellectual property rights could be infringed.

To prevent the abuse of digital content, a number of approaches have been proposed. Digital watermarking is the most widely used form of copy protection. A digital watermark, which is an identification code that carries information about the copyright owner, is invisible and permanently embedded in digital data for copyright protection, proof of ownership, and integrity checks of digital content. It can also provide evidence of copyright infringement. Though useful, watermark-based protection systems have some significant limitations. First, watermarking could degrade the quality of digital content. Second, embedded watermarks are not expected to survive

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 415 – 425, 2006.

under several kinds of attack. In practice, although many techniques have been proposed, watermark-based techniques are not robust enough to prevent malicious users removing watermark via post-processing.

The Digital Rights Management (DRM) system is another popular method for protecting high-value digital assets. DRM is a protocol of hardware and software services and technologies that governs the authorized use of digital content and manages its use throughout the entire life-cycle of the content (as defined by IDC [3]). The primary objective of DRM is to build a DRE (digital right enforcement) environment that only allows access to protected content under the conditions specified by the content owner. Many DRM and DRE frameworks [3][14][16][17] have been proposed in recent years. Although these architectures provide a way to construct a copyright protection environment, the security of digital content is not fully addressed. For example, in the area of rights enforcement, authorized users could still distribute digital assets easily after they pass the identity authentication process. Hence, how to enforce the usage rules and protect content owners' property rights after digital content have been released is still a challenging aspect of DRM research.

Recently, the concept of content-based copy detection has been proposed as a complementary solution for traditional DRM systems. The idea is that, instead of hiding additional information in the digital content (such as digital images and videos) for copy detection, the content itself can be employed. A content-based copy detection system works as follows. It starts by extracting features from the original content, and compare them with features extracted from a suspicious to determine whether the latter is a copy. Content-based copy detection itself can be used to identify illegal copies, or it can be used to complement digital watermarking techniques.

In the past, existed content-based image copy detection techniques [4][11][13] emphasize on finding unique image features with good performance that could resist a variety of image attacks, but finding a globally effective feature is difficult, and in many situations, domain dependent. Hence, the accuracy of image copy detectors is still restricted.

With respect to video copy detection, most approaches [7][8][12][19] employ high-cost computation techniques to match videos, whereby a fix-sized window that slides frame by frame is used to detect copies. However, the window cannot handle some temporal variations, e.g., fast and slow motion. These drawbacks inevitably impede the practicability of the system.

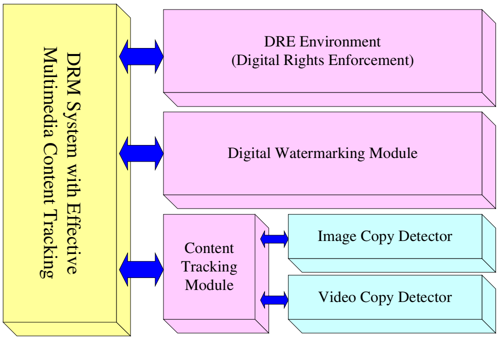

DRM System with Effective
Multimedia Content Tracking
DRE Environment
(Digital Rights Enforcement)
Digital Watermarking Module
Content
Tracking
Module
Image Copy Detector
Video Copy Detector

Fig. 1. Overview of the proposed system

As current copy protection technologies have certain limitation, in this paper, we seek to address the problem by introducing a novel architecture that integrates a DRM system with an effective content tracking mechanism to discourage attackers and further strengthen the proposed system's security. The remainder of this paper is organized as follows. Section 2 introduces our main framework. Sections 3 and 4 describe the proposed content-based copy detection methods in detail. In Section 5, we present the experiment results and demonstrate the effectiveness of content tracking mechanism. Our conclusions are presented in Section 6.

2 Effective Content Tracking for DRM system

Most of the valuable digital content to be protected in archival systems consists of multimedia objects, such as digital images and videos. Figure 1 gives an overview of the proposed DRM system, which consists of three building blocks: (1) the DRE (Digital Rights Enforcement) Environment, (2) the Digital Watermark Module, and (3) the Content Tracking Module. First, the system packages the content to be protected in a secure manner, and the DRE environment ensures that the usage rules are enforced. We use a wrapper-based DRE technique [10] to protect the digital rights. When a user downloads digital content from the network and views it on a player (e.g., a browser), the wrapper automatically monitors the user's behavior. If the rules are violated, or the user refuses to be monitored by the wrapper, the content is rendered unavailable. The second component, the digital watermarking module, can embed an invisible digital watermark into digital content. If necessary, the content holder can extract the watermark to prove ownership if there is copyright infringement. The third component, the content tracking module, can be regarded as the second line of defense. It is composed of two key kernels: an image copy detector and a video copy detector, which can determine whether or not suspicious digital content is copyrighted (registered). By integrating a web crawler with the content tracking module, illegal use of digital content on the Internet can be detected automatically.

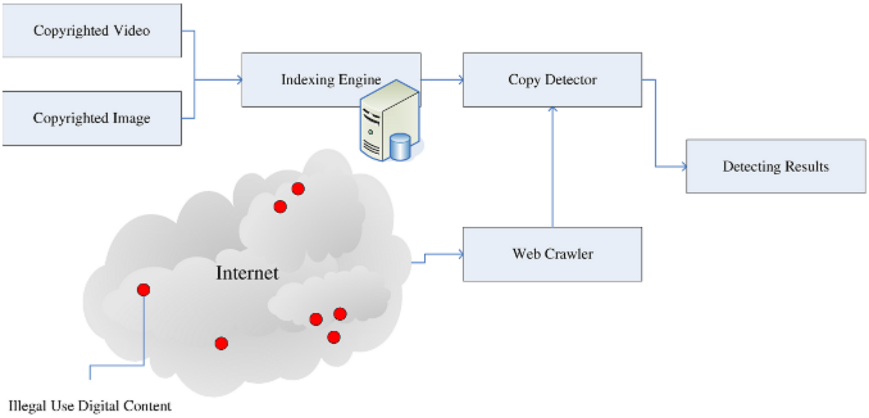


Fig. 2. Block diagram of copy detection

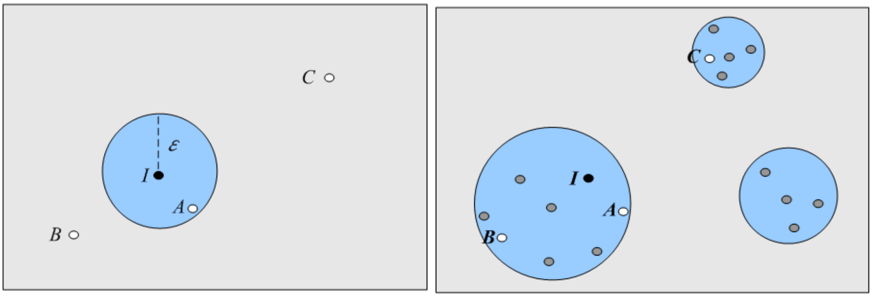
As shown in Figure 2, the content module first registers the image/video with the database. Only feature vectors are stored in the database in order to accelerate the detection process and reduce the amount of storage space required. The image/video copy detector then conducts a matching process to determine whether the suspicious digital content grabbed by web crawler is copyrighted.

3 Image Copy Detection

Previous researchers have tried to find an image feature that can be employed universally for copy detection. Various features have been studied, for example, local [1][18], global [2][11], DCT-based [2], wavelet-based [4][13], geometrically variant [2][4], and invariant [1][13][18]. Obviously, the accuracy of existing copy detectors relies heavily on the robustness of the feature used, and on a suitable threshold that can balance the false rejection and false acceptance rates. However, as we know that it is difficult to find a unique feature that is invariant to various kinds of attack. Another limitation of existing approaches is that they lack a mechanism to exploit useful priori information, such as possible attack models, to boost the copy detection performance – even when such information is easy to generate or acquire.

Hence, instead of extracting the feature vector from a copyrighted image, we use virtual attacks as prior guidance to conduct a new copy-detection framework[9]. Typical attacks considered in our approach include signal-processing attacks, geometric attacks, and image-compression attacks. By applying the attacks to a copyrighted image, a set of novel images can be generated. Both the copyrighted and novel images are processed by extracting their features, where the features extracted from the former and the latter are referred to as the *original* and *extended features* in our framework, respectively. Figure 3 shows the concept of copy detection in a 2-dimensional space. In Figure 3(a), I denotes the feature vector of a copyrighted image, and A, B, and C are the copyrighted images under some malicious attacks. The radius of cluster ε denotes the error tolerance for copy detection in the feature space, which is decided by a predefined threshold. It often occurs that some attack, say A, can be successfully resisted, but the others more severe ones B and C cannot be detected since they are far away from A in the feature space. In our experience, this problem is difficult to solve in practice by simply changing the features being used. Figure 3(b) shows the concept of using *EFS* (extended feature set) to enhance the performance of copy detection, where the gray points denote the extended features. In this case, the problem can be solved by grouping features so that the modified images A, B, and C can be identified correctly.

Although modeling copy detection as a one-class classification problem is likely to boost the system's performance, many empirical studies of pattern classification reveal that the classifier can be trained better if much more prior knowledge is given. In particular, if some negative examples are available, using them would help build a better classifier than using only positive examples. Therefore, in our approach, not only positive examples (where they are mainly extended features), but also negative examples are used. The negative examples are easy to acquire or generate; for example, they can be obtained from the Internet. Also, a registered image can serve as a negative example of another registered image. Our framework transforms the



(a) (b)

Fig. 3. (a) A typical image copy detection algorithm. (b) Using EFS to solve the problem in (a).

copy-detection problem into a two-class classification problem. We demonstrate by experiments that our approach generally outperforms the conventional technique when the same feature space is employed.

A popular method for solving the two-classification problem is based on GMM (Gaussian Mixture Model), defined as:

$$f_k(x \mid \theta) = \sum_{j=1}^{k} w_j g(x \mid \lambda_j),$$

where $g(x|\lambda_j)$ is a multivariate Gaussian distribution, $\lambda = (u,\Sigma)$ is the Gaussian component parameter set, w_j is the weight of jth component, k is the number of Gaussian components, and $\theta=\{w_j, \lambda_j \mid j = 1,2,\ldots,k\}$ is the model's parameter set.

To learn the GMM model for each class, we apply the expectation-maximization (EM) algorithm that can converge to a maximum likelihood estimation of the parameter set. The selection cluster number k is a critical factor in training a GMM [6]. Since we have prior categorical knowledge about our training data, the number of clusters can be set, in advance, as the number of attacks we would like to model. To improve the accuracy, k can also be assigned automatically by maximizing the log-likelihood of the training samples, and estimated via cross-validation. In our approach, we initially set k as the category number, and continue adding clusters until the log-likelihood either (1) starts to decline or (2) keep on increasing but with an amount less than a specific threshold. In Section 5, we conduct some experiments to examine the performance of the proposed framework when a Gaussian mixture classifier is used.

4 Video Copy Detection

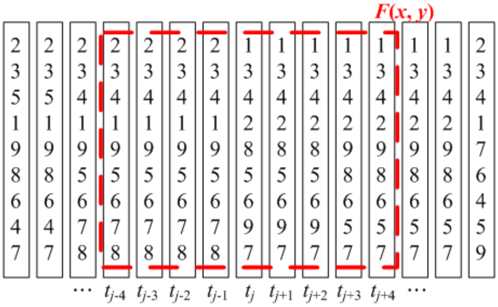
The problem definition of video copy detection is to determine if a given video clip (query) appears in another video clip (target) which is doubtful. However, if it does appear, we need to determine its location. The proposed video copy detection module is responsible for three steps: key frame extraction, candidate clip selection, and sequence matching. Suppose that Q_V and T_V are the query and target video clips, respectively. Q_V is represented as $\{q_{vj} \mid j = 1, 2, \ldots, N\}$, and T_V as $\{t_{vi} \mid i = 1, 2, \ldots, M\}$, where M and N are the number of frames, $M >> N$ and t_{vi} and q_{vj} are the ordinal signatures of the corresponding frames. The details of the ordinal signature are as follows: A video frame is partitioned into $n_x \times n_y$ blocks and the average luminance level in each block is computed. In our case, we utilize 3×3 block of each frame for ordinal signature extraction. Then we sort the set of average intensities in ascending order and a rank is assigned to each block. The ranked $n_x \times n_y$ dimensional sequence is then generated [7][8][12][19]. Thus a video frame is partitioned into 3×3 blocks , as its ordinal signature a 3×3 matrix. We then reshape the matrix to a 9×1 vector. Based on the steps mentioned above, the task of copy detection is to find the subsequences from T_V, whose signature series are similar to those of Q_V .

The first step is to extract key frames from video clips. In addition to reduce the storage and computation costs, it can moderate the effects of temporal variations. Let us take the target clip T_V as an example. In order to search the peak or foot of a sequence, we define a 9×9 Laplacian of a Gaussian filter F, which is often used to calculate second order derivatives in a signal:

$$F(x, y) = -\frac{1}{\pi\sigma^4}\left|1-\frac{x^2+y^2}{2\sigma^2}\right| e^{\frac{x^2+y^2}{2\sigma^2}} ,$$

The second order derivatives reveal signal transitions, which can be chosen as key frames.

We then convolute F and T_V to obtain a vector A and find the local extreme on A, as shown in Figure 4. The extracted key frames are denoted as $T_K = \{t_{k1}, t_{k2}, \dots , t_{km}\}$. For the query clip Q_V, we repeat the above procedure to extract Q_V's key frame sequence $Q_K = \{q_{k1}, q_{k2}, \dots , q_{kn}\}$.



		…	t_{j-4}	t_{j-3}	t_{j-2}	t_{j-1}	t_j	t_{j+1}	t_{j+2}	t_{j+3}	t_{j+4}	…		
2	2	2	2	2	2	2	1	1	1	1	1	1	1	2
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
5	5	4	4	4	4	4	4	4	4	4	4	4	4	4
1	1	1	1	1	1	1	2	2	2	2	2	2	2	1
9	9	9	9	9	9	9	8	8	8	9	9	9	9	7
8	8	5	5	5	5	5	5	5	5	8	8	8	8	6
6	6	6	6	6	6	6	6	6	6	6	6	6	6	4
4	4	7	7	7	7	7	9	9	9	5	5	5	5	5
7	7	8	8	8	8	8	7	7	7	7	7	7	7	9

Fig. 4. The convolution of the filter F and the target T. The dash square indicates the range of F, and t_j is the ordinal signature of the j-th frame in T.

After the key frames has been extracted, the key frame sequence of T_K is still very long. To avoid an exhaustive search of the long sequence, we roughly scan T_K to find subsequences that may be copies of Q_K. First we search for the start and end indices of candidates CI_{start} and CI_{end} in TK. These candidates are frames that are similar to the first and end frames of Q_K (i.e., q_{k1}). Then we scan the second candidate lists CI_{start} and CI_{end}. A subsequence $C = \{t_{ks}, t_{ks+1}, \dots , t_{ke}\}$ in T_K is reported as a candidate clip according to following conditions: First, keep the order of the start and end candidates. Second, select the smallest frame set from the candidate combinations. Third, filter out clips that are too long or too short.

Finally, the sequence matching be processed to compute which clip is similar as copy, hence the Dynamic Time Warping (DTW) algorithm is applied to compute the similarity between the query example Q_K and the candidate clip C. Since DTW can compensate for differences in length, it is suitable for dealing with video temporal variations in videos. We define the following distance function:

$$dist(Q_K, C) = cost(n, l),$$

where n and l are the frame number of Q_K and C respectively, and $cost(n, l)$ is a recursive function:

$$cost(1,1) = \|qk_1 - tk_1\|,$$
$$cost(n,l) = \|qk_n - tk_l\| + min\{cost(n-1,l), cost(n,l-1), cost(n-1,l-1)\},$$

where $\|l - n\| \le$ the maximum warping distance, which is normalized to determine whether Y is a copy of Q_K.

5 Experiment Results

In this section, we conduct two experiments for evaluating the performance of content-based copy detection. We divide the experiments into two cases. In the first case, the detection results of image content tracking are presented; while in the second, video data from the National Digital Architecture Program in Taiwan is used to verify the effectiveness of our method.

5.1 Image Detection Results

We took Kim's approach – DCT ordinal measure [11] as the basis for comparison. In this approach, an input image is divided into 8×8 equal-sized sub-images. Only AC coefficients of the 8×8 DCT coefficients are used as the ordinal measure. We thus generated a 63-dimensional image feature vector.

In the first test, one hundred copyrighted images were registered in the database and used as queries to determine how many modified versions could be detected successfully. A standard benchmark, Stirmark 4.0 [15], was used to generate novel testing data. The image replicas were randomly generated by StirMark 4.0 with 7 categories of pre-learned image attacks (convolution filtering; cropping; JPEG; median filtering; noise adding; scaling; and rotation), and 6 categories of novel attacks, including affine transformation, self-similarity, removal of lines, PSNR, rotation+rescaling (abbreviated as RRS), and rotation+cropping (abbreviated as RC). We also generated 124 near-replicas for each copyrighted image. In addition to image replicas, the testing data also contained 15,000 randomly picked unrelated images, giving a total of 27,400 images for testing.

To evaluate the performance, the precision rate, recall rate, and F-measure are used:

$$F - Measure = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}.$$

The results in Table 1 show that our framework outperforms that of the DCT ordinal measure. The EFS for the Gaussian mixture model achieve very high precision and recall rate of 96.56% and 93.54% respectively, while the F-measure is 95.03.

Table 1. Average precision and recall rates by using extended features and pure DCT ordinal measures. The response time consists of both the feature extraction and classification times.

Algorithm	Avg. Precision	Avg. Recall	F-Measure
DCT ordinal measures	93.13%	54.79%	68.99
Gaussian Mixture Classifier with EFS	96.56%	93.54%	95.03

Fig. 5. Three color images in the digital museum (512*512 pixels): a container, a rare book and a painting were chosen in the second experiment

Table 2. Recognition rates of the Gaussian Mixture Classifier (including novel attacks): The first column indicates whether the type of image attack was pre-learned, while the second column shows the attack model and how many times it was applied. For example, " Noise * 12" means that the noise attack was applied to the image 12 times. In the other columns, "m(n)" indicates that the number of image replicas successfully detected by our Gaussian mixture method and by the pure DCT ordinal measures method was m and n respectively.

Pre-learned	Testing Item	Container	Rare Book	Painting
∨	Convolution Filtering * 2	2(2)	2(2)	2(2)
∨	JPEG * 14	14(14)	14(14)	14(14)
∨	Median Filtering * 4	4(4)	4(4)	4(4)
∨	Noise * 12	12(10)	12(9)	12(8)
	Self-Similarities * 3	3(3)	3(3)	3(3)
	PSNR * 10	10(10)	10(10)	10(10)
∨	Scaling * 10	10(10)	10(10)	10(10)
∨	Cropping * 13	9(1)	8(2)	11(0)
∨	Rotation * 18	17(0)	16(0)	18(0)
	Affine * 8	7(7)	8(6)	7(6)
	Removing Lines * 10	10(8)	10(8)	10(8)
	RRS * 10	9(1)	9(1)	9(0)
	RC * 10	9(1)	9(1)	9(0)
Recognition Rate (DCT ordinal measures)		(71+70+65) / (124*3) = 55.38%		
Recognition Rate (Gaussian Mixture Classifier with EFS)		(116+115+119) / (124*3) = 94.07%		

The above experiment shows the overall performance of our method. To test the robustness against different attacks, we conducted another smaller-scale experiment in which only the three images shown in Figure 5 were used. This allowed us to show the comparisons of the performance of EFS with conventional copy detection method in more detail. The results are summarized in Table 2. We also applied some novel attacks (i.e., attacks not modeled in the training phase) to examine the performance of our approach. The results show that the images' resistance to geometric attacks (cropping, rotation, scaling) was significantly enhanced by our approach; on average, more than half the manipulated geometric images were correctly identified in the experiment. In Table 2, the first column indicates whether the image attacks were prelearned. Clearly, for those novel attacks we did not model in advance, our approach still achieves an acceptable performance and outperforms the pure DCT ordinal measure method.

5.2 Video Detection Results

We experimented with approximately 106,333 frames of video data from the NDAP's digital video library of social culture in Taiwan. The format of the videos is MPEG-1 NTSC, for which the resolution is 352×240 and frame rate is 29.97 fps. To test the performance of the proposed approach, the video data was modified to generate eight copies for brightness, histogram equalization, changing the resolution to 176×120, changing the frame rate to 15 fps and 10 fps, slow motion (0.5×), fast motion (2×), and hybrid modification (changing to 176×120 resolution, 10 fps, and 2× fast motion). We randomly selected 100 video clips (100×1000 frames in total) as query clips for each type of copy. Hence there are 800 queries in the experiment to verify the track performance in our video copy detection module.

Table 3. The F-measure of brightness, equalization, and frame size changing (spatial variations), and frame rate changing, slow and fast motion (temporal variations) copy in Hua's, Kim's and our proposed approach

	Brightness	**Equalization**	**176×120**	**10fps**	**15fps**	**0.5×**	**2×**	**Hybrid**
Hua	89.98	94.87	90.13	94.25	96.01	53.27	75.94	65.52
Kim	93.61	95.89	93.14	76.54	85.90	25.86	43.30	40.27
Proposed	94.26	96.19	94.24	93.87	95.60	83.55	94.06	83.38

We compared the results of the proposed approach with Hua's [7] and Kim's [11] approaches using F-measure. Table 3 shows the F-measure of all cases, and our approach outperforms the other two greatly. According to the experiment results, we see that our method performs slightly better than Hua's and Kim's for spatial variation attacks such as brightness, equalization and frame size change. For frame rate changes, our method performs better than Kim's but slight worse than Hua's. However, our method achieves a far better performance (average F-measure is 88.81) for the attacks of fast and slow motion than those of the others (average F-measure is 64.61 and 34.58). To conclude, our method has better performance in overall for the hybrid case, and is effective not only for spatial-variation but also temporal-variation attacks.

6 Conclusions

Protecting digital content presents serious technical challenges that the existing approaches have not overcome. The integrated framework presented in this paper provides a solution for digital content protection of digital libraries. With the wrapper-based DRE technique, a digital rights enforcement environment can be built to maintain the usage rules of digital content. With the help of such content tracking mechanism, pirated digital content altered from original images and videos can be effectively identified. Also, the introduced copy detection techniques have been demonstrated to be more accurate than traditional approaches. By employing such a complementary design, the abuse of valuable digital content can be prevented, and further reduce the copyright infringements.

Acknowledgement

This work was partially supported by National Science Council, Taiwan, under Grants NSC94-2422-H-001-006, NSC94-2422-H-001-007, NSC95-2422-H-001-007, and NSC95-2422-H-001-008.

References

1. S.-A. Berrani, L. Amsaleg, and P. Gros, "Robust content-based image searches for copyright protection," *Proceedings of the 1st ACM international workshop on Multimedia databases*, pp. 70-77, 2003.
2. D.-N. Bhat and S.-K. Nayar, "Ordinal measures for image correspondence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 20 , Issue 4, pp. 415-423, 1998.
3. C.P. Bogdan et al., "A DRM security architecture for home networks," *Proceedings of the 4th ACM workshop on Digital rights management*, pp 1-10, 2004.
4. E.-Y. Chang, J.-Z. Wang, C. Li, and G.. Wiederhold, "RIME: a replicated image detector for the world-wide-web," *Proceedings of the SPIE Multimedia Storage and Archiving Systems*, San Jose, CA, November 1998.
5. J. Duhl, S. Kevorkian, "Understanding DRM Systems," *An IDC Research White Paper*, 2001.
6. M.A.F. Figueiredo and A.-K. Jain, "Unsupervised learning of Finite mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume: 24 Issue: 3, pp. 381–396, March 2002.
7. A. Hampapur, K.-H. Hyun, and R. M. Bolle, "Comparison of sequence matching techniques for video copy detection," *The SPIE Conference on Storage and Retrieval for Media Databases*, pp. 194-201, 2002.
8. X. S. Hua, X. Chen, H. J. Zhang, "Robust video signature based on ordinal measure," *International Conference on Image Processing*, 2004.
9. J.-H. Hsiao, C.-S. Chen, L.-F. Chien, and M.-S. Chen, "Image Copy Detection via Grouping in Feature Space Based on Virtual Prior Attacks," *Proceeding of International Conference on Image Processing*, Atlanta, GA, USA, 2006.

10. J.-H. Hsiao, J.-H. Wang, M.-S. Chen, C.-S. Chen and L.-F. Chien, "Constructing a Wrapper-Based DRM System for Digital Content Protection in Digital Libraries," *Proceedings of the 8th International Conference on Asian Digital Libraries*, ICADL 2005, pp. 375-379, 2005.
11. C. Kim, "Content-based Image Copy Detection," *Signal Processing: Image Communication*, Vol 18, pp. 169-184, 2003.
12. C. Kim and B. Vasudev, "Spatiotemporal sequence matching for efficient video copy detection," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 15, No. 1, pp. 127-132, 2005
13. C.-S. Lu, C.-Y. Hsu, S.-W. Sun, and P.-C. Chang, "Robust mesh-based hashing for copy detection and tracing of images," *Proceedings of IEEE International Conference on Multimedia and Expo*, Volume 1, pp. 731-734, June 2004.
14. T. Nicolakis, C.E. Pizano, B. Prumo and M. Webb, "Protecting Digital Archives at the Greek Orthodox Archdiocese of America," *Proceedings of the 2003 ACM workshop on Digital rights management*, October 2003.
15. F. Petitcolas, "Watermarking Schemes Evaluation," *IEEE Signal Processing Magazine*, Volume 17, No. 5, pp. 58-64, 2000.
16. A.J. Pramod and L.H. Gregory, "DRM as a layered system," *Proceedings of the 4th ACM workshop on Digital rights management*, pp 11-21, 2004.
17. G. Susanne, "A Sample DRM System," *Lecture Notes in Computer Science*, Volume 2770, pp. 150-161, 2003.
18. K. Yan, R. Sukthankar, and L. Huston, "Efficient Near-duplicate Detection and Subimage Retrieval," *Proceedings of the 12th annual ACM international conference on Multimedia*, session 13, pp. 869-876, 2004.
19. J. Yuan, Q. Tian, and S. Ranganath, "Fast and robust search method for short video clips from large video collection," *International Conference on Pattern Recognition*, Cambridge, UK, 2004.

Semantic Web Techniques for Multiple Views on Heterogeneous Collections: A Case Study

Marjolein van Gendt[1,2], Antoine Isaac[1,2,3], Lourens van der Meij[1,2], and Stefan Schlobach[1]

[1] Vrije Universiteit Amsterdam
[2] Koninklijke Bibliotheek, Den Haag
[3] Max Planck Institute for Psycholinguistics, Nijmegen
{mtvgendt, aisaac, lourens, schlobac}@few.vu.nl

Abstract. Integrated digital access to multiple collections is a prominent issue for many Cultural Heritage institutions. The metadata describing diverse collections must be interoperable, which requires aligning the controlled vocabularies that are used to annotate objects from these collections. In this paper, we present an experiment where we match the vocabularies of two collections by applying the Knowledge Representation techniques established in recent Semantic Web research. We discuss the steps that are required for such matching, namely formalising the initial resources using Semantic Web languages, and running ontology mapping tools on the resulting representations. In addition, we present a prototype that enables the user to browse the two collections using the obtained alignment while still providing her with the original vocabulary structures.

1 Introduction

Integrated access to multiple digital collections is a prominent issue within many research departments of Cultural Heritage (CH) institutions. These collections contain different kinds of objects, with different subjects, are described using different annotation schemes and controlled vocabularies and might be stored in and be accessible via different information systems: they are heterogeneous.

To access several such sources via one portal, one first needs to obtain syntactic interoperability by building a system that can get information from all sources simultaneously, using standard protocols or shared metadata schemes.

However, to maximally use the original resources, integrated systems should also tackle the hitherto unsolved *semantic interoperability* problem, i.e. properly take into account the conceptual similarities and differences between collections. Linking subject descriptors from the vocabularies used to annotate the different collections (e.g. between "birds" in one vocabulary and "flying beings" in another one) provides such interoperability. But it also necessary to keep the original semantics of these vocabularies, such as found in the hierarchical relations between descriptors (e.g. "birds" as specialization of "animals").

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 426–437, 2006.

The general aim of our project, STITCH[1], is to determine to what extent *Semantic Web* (SW) techniques, such as ontology alignment, can solve these interoperability issues. As CH vocabularies are similar to ontologies, adapting this research to the CH sector seems promising.

Our first experiment and implementation aimed at providing integrated access to two heterogeneous collections, the Illuminated Manuscript collection[2] from the Dutch National Library (KB), and the ARIA Masterpieces collection[3] from the Rijksmuseum in Amsterdam. In this paper, we describe the concrete steps of this experiment. First, a conversion to generic formats, such as RDF(S)[4] and SKOS[5], was required to provide integrated access to semantically linked CH collections. Second, we could align them using these computer-readable representations. We turned to two off-the-shelf ontology mappers (S-Match [8] and Falcon [11]) and evaluated their use for aligning CH controlled and structured vocabularies. Third, automatically found correspondences were used in a purpose-built interface for browsing different vocabularies and retrieving documents from several collections in parallel, based on the multi-faceted browsing paradigm.

As said, the goal of the research described in this paper is to evaluate the potential and limits of current Semantic Web technology for integrating multiple CH collections with heterogeneous vocabularies. Our main research questions are:

1. Are the current SW techniques suitable for solving this integration problem?
2. Are there specific CH problems that need particular efforts from the SW community?

The paper is structured as follows. In Section 2 we introduce our case study, by describing the two collections we aligned. In Section 3 we describe our solution to the problem from a practical perspective. In Section 4 we then discuss the relevance of our findings for both CH and SW practitioners, before we relate our work to existing work, and conclude.

2 Case Study: Illuminated Manuscripts and Masterpieces

The Illuminated Manuscripts and Masterpieces collections contain objects such as images, drawings, books and/or sculptures. Most interesting for us is the heterogeneity of the vocabularies used to describe these collections.

The Manuscripts collection contains 10.000 medieval illuminations which are, in addition to the standard bibliographical information, annotated by subject indices describing the content of the image. These indices come from the Iconclass classification scheme, a 25.000 element vocabulary with iconographical analysis as main purpose. An Iconclass *subject* consists of a *notation* – an alphanumeric identifier used for annotation – and a *textual correlate* – e.g. `25F9`

[1] SemanTic Interoperability To access Cultural Heritage, `http://stitch.cs.vu.nl`

[2] `http://www.kb.nl/kb/manuscripts/`

[3] `http://www.rijksmuseum.nl/collectie/index.jsp?lang=en`

[4] `http://www.w3.org/RDF/`

[5] `http://www.w3.org/2004/02/skos/`

`mis-shapen animals; monsters`. Subjects are organized in nine hierarchical trees. Other features are associative *cross-reference* links as well as mechanisms for subject specialisation, such as *keys* – e.g. `25F9(+33)` would refer to the head of a monster. Additionally, subjects have simple *keywords* used for retrieving them: `25F9` is thus linked to "`monster`" and "`shape`", amongst others. It is important to note that textual correlates are often in the form of glosses, e.g. `Noah's sacrifice; various animals are offered, possibly a lamb, a dove and a ram (often combined with the rainbow of the covenant)`.

The Masterpieces collection contains 700 objects such as paintings and sculptures and its subjects are indexed using the ARIA "catalogue". This controlled vocabulary, conceived mainly as a resource for browsing, consists of about 500 terms and three sub-vocabularies. The first is intended for the layman, and contains subjects like `Man`, while the second is for more advanced users: it contains similar but finer-grained subjects like `Male portraits`. A third very small list – 6 types of objects, like `Sculpture` – is used as a high-level entry point to the system. The only "semantic" information found in this catalogue consists of specialisation links within the first two vocabularies, that can be interpreted as classical "Broader Than" relationships. The hierarchies are only two levels deep and there are occurrences of multiple inheritance.

3 Performing the Case Study

In this section we describe our approach for providing access to the integrated Illuminated Manuscripts and ARIA Masterpieces collections. Figure 1 shows our framework in a schematic way. In a first step we transform both collections and their respective thesauri into Semantic Web compliant representation languages. Secondly, we create an alignment between the two thesauri using existing mapping technology. Finally, we build a browser to access the linked collections.

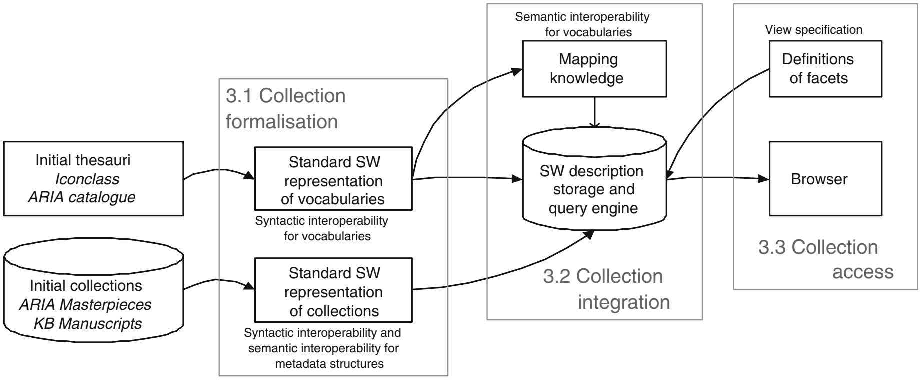

Initial thesauri
Iconclass
ARIA catalogue
Initial collections
ARIA Masterpieces
KB Manuscripts
3.1 Collection formalisation
Standard SW representation of vocabularies
Syntactic interoperability for vocabularies
Standard SW representation of collections
Syntactic interoperability and semantic interoperability for metadata structures
Semantic interoperability for vocabularies
Mapping knowledge
SW description storage and query engine
3.2 Collection integration
View specification
Definitions of facets
Browser
3.3 Collection access

Fig. 1. The different steps of our experiment

3.1 Collection Formalisation

This case study supplies two types of CH resources that need transformation: the controlled vocabularies and the collections themselves.

Converting controlled vocabularies. There have been substantial methodological efforts concerning the conversion of CH vocabularies into SW formats. Similar to [1], we handle the knowledge acquisition process in two steps: first, analysing the sources our use-case provided, and second, formalising the knowledge they contain. This last step involves two consecutive conversions, to first get a standard representation and then an application-specific one.

Analysis. We had the controlled vocabularies as well as significant expert feedback at our disposal. As the vocabularies differ significantly in nature and use, we expected them to be difficult to represent using the same formal apparatus. The question was whether to take all peculiarities of the respective vocabularies into account, or to turn to some standard model. We opted for the latter, as we wanted to test a process – both for representation, alignment and exploitation – that could be generalized to a wider range of vocabularies.

Standard formalisation. The SKOS (Simple Knowledge Organisation System) initiative provides a standardized model to encode the most common knowledge organization schemes, such as thesauri or classification schemes, in SW languages. SKOS is an RDF vocabulary that is currently being developed within the W3C Semantic Web activity. ARIA proved almost fully compatible with the SKOS schema. We only managed to convert Iconclass subjects partly: SKOS could not cope with Iconclass idiomatic elements, such as keys.

Application-specific formalisation. Tools such as storage engines or browsers should interpret the SKOS files in accordance with their intended semantics. This often requires tweaking, e.g. to make our generic RDFS engine deal with the transitivity of the SKOS `broader` relation we had to interpret it as a sub-property of RDFS `subClassOf`.

Converting collection elements. Our main focus being description *vocabularies*, we just used the description *structures* as they were in the original collections, without enforcing a unified scheme like Dublin Core. From the two metadata schemes we constructed small metadata ontologies in RDF Schema.

3.2 Collection Integration

Having formalised our CH vocabularies in SW-compliant representations has the advantage that we can use existing ontology mapping tools to align them. We applied two state-of-the-art ontology mappers, Falcon and S-Match.

Falcon [11] is one of the best performing tools[6] for aligning complex RDFS/OWL ontologies. It relies on a combination of lexical comparison and graph-matching techniques. First, it compares concepts based on the set of weighted terms derived from their lexical "environment": their own identifiers, labels, comments, but also the ones of their immediate neighbors – parents, children – in the ontology. These similarities are used as input for the second step, which exploits a graph representation of the semantic information and matrix computation processes to finally return equivalence links between the concepts and relations of the compared ontologies.

S-Match. [8] has been developed for mapping vocabularies represented as trees. It has a modular approach where a *lexical* matching component, a background-knowledge component ("*oracle*") and a *structure-based* mapping module all contribute to computing a mapping between the input trees. In S-Match default configuration, Wordnet[7] is used as the background knowledge component.

S-Match is not a general ontology mapper, but specializes on hierarchical classification trees used to structure the access to documents. S-Match core mapping method exploits the fact that the meaning of a concept in such a tree is determined by the concepts in the path to the root. Based on the lexical component and the oracle, each concept is associated with a propositional formula representing all its "available meaning". The mapping relations are then determined by the logical relations between the formulas for the concepts of the to-be-aligned classification trees.

Mapping results. In table 1 some good mappings produced by S-Match are shown, where the first mapping was produced mainly based on lexical mapping, the second using stemming, and the third making use of background knowledge.

Table 1. Some good S-Match mapping results

IC notation	Iconclass textual correlate	Relation	ARIA label
23L	'the twelve months represented by landscapes'	Less General	'Landscapes'
25A271	'(map of) the North Pole'	Less General	'Charts, maps'
23U1	'calendar, almanac'	Less General	'Publications'

Mapping thesauri proved to be difficult for both mappers, and the overall results were less than satisfactory. Evaluation measures for mapping results depend on their intended use. Regarding our intended browsing interface, precision is more important than recall, because we do not want to confront users with useless links. For S-Match a precision of 46% is obtained on a selected subset of Iconclass (1500 concepts) and the complete ARIA thesaurus (500 concepts); 46% of the mappings were correct. Falcon reached a precision of only 16%.

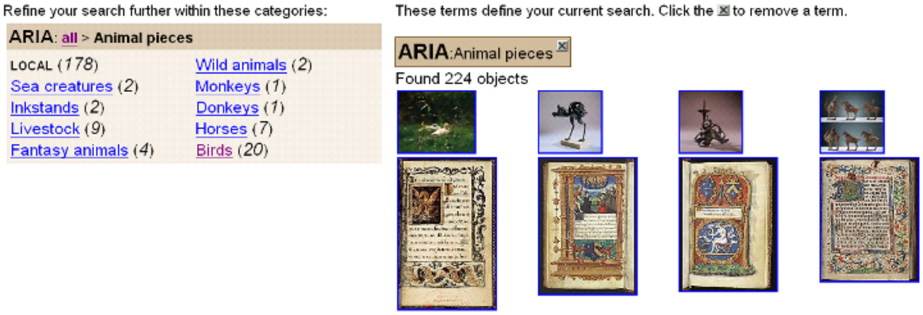
[6] See the 2005 OAEI campaign, `http://oaei.ontologymatching.org/`

[7] `http://wordnet.princeton.edu/`

3.3 Collection Access

We implemented a multi-faceted browsing (MFB) framework to evaluate and explore the results of our mapping effort. MFB involves constraining search criteria along – usually orthogonal – aspects of a collection called *Facets*. Here we tuned the MFB paradigm in an atypical way, since we used one category (the *subject* annotation) for defining several facets. Such a setting is possible because objects are often annotated by several subjects. So using one facet to search for "monkey" and another for "landscape" could retrieve pictures of a monkey in a landscape.

For searching through the integrated collections we explored three different views on integrated collections: *single*, *combined*, and *merged view*.



Refine your search further within these categories:

ARIA: all > Animal pieces

LOCAL (178)
Sea creatures (2)
Inkstands (2)
Livestock (9)
Fantasy animals (4)
Wild animals (2)
Monkeys (1)
Donkeys (1)
Horses (7)
Birds (20)

These terms define your current search. Click the ☒ to remove a term.

ARIA:Animal pieces ☒

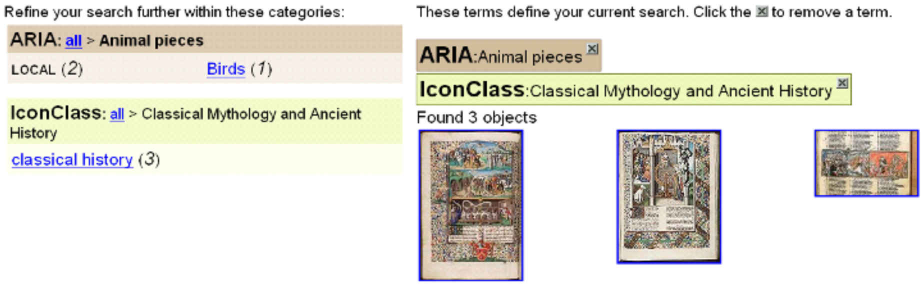
Found 224 objects

Fig. 2. Single View: Using the ARIA thesaurus to browse the integrated collections

The *Single View* presents the integrated collections from the perspective of only one of the collections. The elements of the other collection are made accessible by means of the correspondences between their subject annotations and the concepts of the current view. In figure 2 the first four pictures come from the Rijksmuseum, the others are Illuminated Manuscripts. Browsing is done solely using the ARIA Catalogue, i.e. these illuminations have been selected thanks to the automatically extracted mapping between ARIA concept "Animal Pieces" and Iconclass "25F:animals".

The *Combined View* provides simultaneous access to the collections through their respective vocabularies in parallel. This allows us to browse through the integrated collections as if it was a single collection indexed against two vocabularies. In figure 3 we made a subject refinement to ARIA "Animal pieces", and narrowed down our search with Iconclass to the subject "Classical Mythology and Ancient History". Only three Manuscripts matched these criteria. Notice that we browse according to a "biological" criterion using ARIA, and a "mythological" one from Iconclass to come to our results.

The *Merged View* provides access to the collections through a merged thesaurus combining both original vocabularies into a single one, based on the links found between them in the automatic mapping process. For figure 4 we made the same selection as for the "single view" case. But notice that the "merged



Refine your search further within these categories:

ARIA: all > Animal pieces

LOCAL (2) Birds (1)

IconClass: all > Classical Mythology and Ancient History

classical history (3)

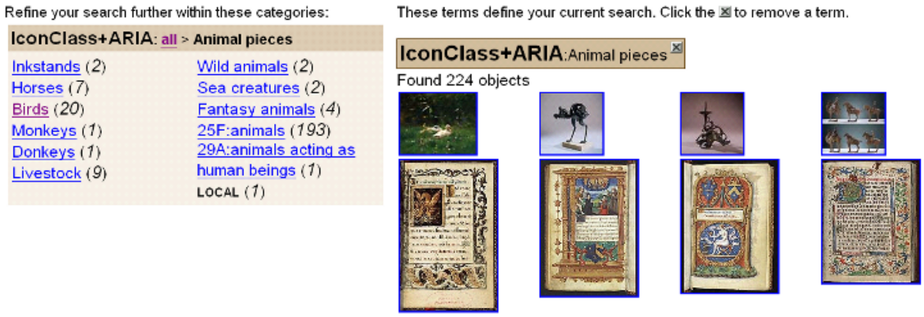
These terms define your current search. Click the ☒ to remove a term.

ARIA:Animal pieces ☒

IconClass:Classical Mythology and Ancient History ☒

Found 3 objects

Fig. 3. Combined View: Using ARIA and Iconclass to browse the integrated collection



Refine your search further within these categories:

IconClass+ARIA: all > Animal pieces

Inkstands (2)	Wild animals (2)
Horses (7)	Sea creatures (2)
Birds (20)	Fantasy animals (4)
Monkeys (1)	25F:animals (193)
Donkeys (1)	29A:animals acting as human beings (1)
Livestock (9)	LOCAL (1)

These terms define your current search. Click the ☒ to remove a term.

IconClass+ARIA:Animal pieces ☒

Found 224 objects

Fig. 4. Merged View: Using a merged thesaurus to browse the integrated collection

view" now provides both ARIA concepts such as "Birds" and an Iconclass concept "29A:animals acting as human beings" for further refining our search. The mapping primitives determine the merging: two concepts that are identified to be equivalent are merged into one new concept, and if the mapping determined that a concept from one scheme is broader than a concept in the other scheme, the second concept is added as a child of the first.

Prototype details. The design of our browser was inspired by the Flamenco search interface framework [9]. It is implemented in SWI-Prolog and uses the Sesame RDF repository[8] for storage and querying.

4 Lessons Learned

The main goal of our research was to find out to what extent SW techniques can solve heterogeneity issues when integrating multiple CH collections.

The general conclusion is positive: in a relatively short time we managed to implement an integrated browsing environment that was built purely on accepted standards for representing data, and which used existing tools for storage,

[8] Available on `http://www.openrdf.org`

querying and mapping. However, there is more to be learned for CH collection managers and developers of SW tools alike. In this section, we first try to answer questions concerning the practical relevance of chosen techniques and tools: to which extent can CH use-cases be successfully addressed by such solutions? We then explore the problems raised by our experiment from the point of view of SW researchers. Is our approach methodologically and technologically sound?

4.1 A Cultural Heritage Perspective

Conversion Process. Implementing a realistic process for going from CH resources to SW-compatible formats was successful, but often non-trivial.

Conversion and standards for CH vocabularies. CH vocabularies often rely on complex models that are non-standard, which can hinder the conversion process. Especially for Iconclass some modeling decisions had to be made. For example, for *notations* we used the SKOS `prefLabel` property to enforce the necessary uniqueness constraint, even though notations like `25F9` definitely miss the lexical flavor to make them proper *terms* e.g. `mis-shapen animals; monsters`. Even worse, some features could not be represented at all, like *keys* or the additional network of *keywords*. Potentially interesting information had to be sacrificed for the sake of generality, which illustrates the trade-offs of using standards.

Ontology mapping vs. thesaurus mapping. For our case study we applied off-the-shelf SW ontology mappers. However, CH controlled vocabularies have features that make them really different from ontologies, e.g. glosses for describing concepts instead of simple terms. Here we describe the repercussions these peculiarities have on alignment quality.

Mapping poorly structured schemes. Most ontology mappers rely on structure-based comparison using ontology semantics: subsumption relations, properties, etc. However, thesauri have less strictly defined semantic relations and their consistency is not always enforced. Because of this and the loss of information in the formalisation step, the only usable structural information present in our thesauri is the broader and narrower term hierarchy.

Falcon heavily exploits structure components usually present in expressively modelled ontologies. An analysis of the few correct results from Falcon shows that the lexical mapping works fine, but that the reliance on graph-based techniques usually contributes negatively to the overall process.

S-Match produces much better mapping results, as it was purpose-built for tree-like structures and uses the extensive lexical background information found in Wordnet. Nevertheless, the influence of the difference of the depth levels in both thesauri has unfortunate consequences: the fact that S-Match uses the full path of a classification tree for the mapping implies that its output almost always consist of specialisation links *from* Iconclass concepts *to* ARIA concepts. For browsing, this is very damaging, as it constrains the way a user can specialize her queries: once she is browsing Iconclass subjects, she cannot find ARIA specialisations anymore.

Gloss features and concept matching. The gloss features of concepts cause two anomalies to occur: 1) natural language meaning of a sentence is not interpreted, and 2) the meaning of single terms is not disambiguated by the remainder of their gloss, and thus interpreted as if denoting concepts on their own.

Table 2. Some bad S-Match results

IC notation	Iconclass textual correlate	Relation	ARIA label
23H	'seasons of the year represented by concepts other than [. . .] landscapes [. . .]'	Less General	'Landscapes'
29D	'natural forms in stones, wood, clouds'	Less General	'Jewellery'

An example of a bad match caused by lack of natural language interpretation is the first mapping in table 2: S-Match does not interpret "other than", which causes `23H` to wrongly match `Landscapes`.

Using Wordnet as background knowledge sometimes also leads to finding irrelevant links based on comparing single words, which could have been disambiguated by the other words found in the glosses. In table 2, 'Jewellery' would legitimately map to precious stones, but the other tokens in `29D` should have provided enough information to disambiguate between the different kinds of stones. An option for improvement would be to focus on smaller but more relevant pieces within Wordnet, e.g. taking only closest siblings into account.

4.2 A Semantic Web Perspective

Generalizability. The *Semantic Web* claims to provide generic solutions. Therefore, the question arises whether it would be easy to reproduce what we did with new collections. Surely, we would benefit from the experience we gained in this case study, and the SW frameworks proved to be flexible enough to cope with different representational choices. But the transformation and mapping process would remain case-study dependent in at least two ways: First, the conversion effort depends on the technical and functional requirements implied by the choice of specific *tools* and *tasks*. Second, both conversion and alignment processes are dependent on the CH *resources*. Take for example the influence of the structure of the vocabularies on the mapping process we discussed in the previous section.

Role of standards. In our approach the role of SKOS was crucial. Such a standard helps to integrate the different components of a framework. It also contributes to improving the extendability of the framework: for example, an additional SKOS-encoded thesaurus could be integrated easily in our tools.

The lack of *de facto* standards for alignment tools was a prominent problem. S-Match takes as input indented trees, which caused an important loss of information. Falcon does better, as it admits expressive standard RDF/OWL ontologies. For output things are even worse: Falcon outputs links in a standardized syntax, but its semantics are unclear. Again, S-Match was less generic, as its output is an ad-hoc non-standard format.

Methodological process guidance. The SW community already got concerned with conversion and deployment of CH vocabularies, and has proposed methodological guidelines. Van Assem et. al. [1], for example, advocate three conversion steps. In the first step, the original vocabulary is translated into an RDFS/OWL model that mirrors the original structure as precise as possible. In the second step, one interprets the model so that intended semantic properties can be explicitly assigned to the RDFS/OWL representation. Finally, one can represent the vocabulary using a standard model like SKOS.

In our experiment we took this process as a guidance, although, focusing on generality and implementation matters, we only applied its last two steps. However, for mapping purposes, the process itself might be questionable. On the one hand, using a standard model only, as described in [2], can help aligning vocabularies: a basic part of the integration process is partly dealt with by conversion. On the other hand, in order to give alignment tools more information for mapping, a conversion step specific to each controlled vocabulary could be beneficial.

Scalability. SW solutions are often criticized for their performing poorly against massive data sets, which are common in the CH world. Indeed, as Falcon uses a complex algorithm, it was practically impossible to have it run on complete Iconclass. Some division had to be done beforehand. However, S-Match performed better: it took five hours to achieve a complete alignment, which is not a problem since our application does not need to compute mappings at runtime.

5 Related Work

Our case study has been influenced by portal projects like The European Library[9] and the Memory of the Netherlands[10]. But these do not use correspondences between vocabularies, though this problem has already been identified in the Digital Library DL field [6]. Some DL projects like MACS [3] or RENARDUS[11] have used mappings, but they relied on manual alignment, costly and possibly imprecise. We wanted to explore the use of automatic alignment of concept schemes, like currently done in the SW community. This community produced a number of dedicated tools [12], sometimes inspired by previous database integration efforts [5][12]. However, automatic alignment methods usually lack concrete experiments that would assess the feasibility of integrating them in deployed applications, even when they explicitly focus on the thesaurus field [4].

Our approach is thus closer to settings like [10] or [7] that try to apply SW techniques to concrete (CH) cases, except for our focus on automatic alignment.

[9] `http://www.theeuropeanlibrary.org`

[10] `http://www.geheugenvannederland.nl`

[11] `http://www.renardus.org`

[12] We could have tried to directly turn to such techniques. But while they naturally focus on the structure of data – as encoded in database schemas – we focus on the semantics of descriptors that come in unstructured subject annotations.

Actually [10] also implements faceted browsing; we both were inspired by the Flamenco framework [9]. We could have tried to re-use these solutions; however, availability problems and our need for flexible experiments with various setups made us decide to build our own prototype.

6 Conclusion

In this paper we have presented a case study aiming at solving the semantic interoperability problem in the context of CH resources, using automatic alignment processes between their vocabularies to avoid heavily labour-intensive and ambiguous manual alignment work.

This study provides interesting insights regarding the use of SW techniques in a CH environment. We have seen that the conversion of vocabularies using standardised formats is possible, and helps their deployment. We have also shown that based on such representations and automatically found mappings, an operational interface for browsing heterogeneous collections in an integrated way *can* be implemented.

If all collections and thesauri were available in standard formats (SKOS, RDF) or when automatic conversion is feasible so that translation steps would not be needed anymore, our framework would provide a very easy way of integrating heterogeneous collections. However, there still are problems to solve before this ideal situation occurs:

- we have to overcome the loss of semantics when translating the thesauri into SW standards, for instance by providing more expressive standards,
- ontology mapping tools should be compliant with the SW standards concerning input and output formats, and
- specifically for CH controlled vocabularies, it would be preferable to have a SKOS standard inference engine instead of an RDF(S) one[13].

Furthermore, all tools (mappers, inference engines) should be scalable for handling the enormous amount of data present in CH.

Concerning the use of ontology mappers for our CH case, we learned that available ontology alignment techniques need to be tuned to be of use for e.g. thesaurus mapping. Most mappers use resources that are absent from thesauri, e.g. properties, and refrain from (properly) using all information found in thesauri, e.g. synonyms. S-Match mapping quality (46%) is a lot higher than Falcon one (16%), but must still be improved to be useful for browsing purposes. Typical features such as gloss descriptions and poor structuring should be taken into account when constructing a thesaurus mapper. So, to perform semantic integration of CH collections the way we envision, automated mapping techniques are indispensable, but should absolutely be adapted.

Finally, our interpretation of Multi-Faceted Browsing provides multiple views or access points for a same set of data. This way users can choose the vocabulary

[13] Note the discrepancy between this point and the first: the use of standards limits the amount of transferable information, but provides generalizability.

they are most comfortable with and thus personalised access is granted. We encourage readers to try our browser at http://stitch.cs.vu.nl/demo.html.

Acknowledgements

STITCH is funded by NWO, the Netherlands Organisation for Scientific Research. We would like to thank KNAW and RKD institutes for Iconclass resources, and Gerda Duijfjes-Vellekoop for helping with these. We also owe the Rijksmuseum and CHIP project people (especially Lloyd Rutledge) for the ARIA material, the researchers from Trento University (especially Pavel Shvaiko) for their help with S-Match, and Henk Matthezing for his constant advice at the KB.

References

1. van Assem, M., Menken, M. R., Schreiber, G. et al.: A Method for Converting Thesauri to RDF/OWL. Int. Semantic Web Conference, Hiroshima, Japan, 2004.
2. van Assem, M., Malaise, V., Miles, A., Schreiber, G.: A Method to Convert Thesauri to SKOS. 3rd European Semantic Web Conference, Budva, Montenegro, 2005.
3. Clavel-Merrin, G.: MACS (Multilingual access to subjects): A Virtual Authority File across Languages. Cataloguing and Classification Quarterly 39 (1/2),2004.
4. Constantopoulos, P., Sintichakis, M.: A Method for Monolingual Thesauri Merging. ACM SIGIR Conference, Philadelphia, USA, 1997.
5. Doan, A. and Halevy, A.: Semantic Integration Research in the Database Community: A Brief Survey. AI Magazine, Special Issue on Semantic Integration, 2005.
6. Doerr, M.: Semantic Problems of Thesaurus Mapping. Journal of Digital Information, 1 (8), 2004.
7. Gasevic, D., Hatala, M.: Searching Web Resources Using Ontology Mapping. K-CAP Workshop on Integrating Ontologies, Banff, Canada, 2005.
8. Giunchiglia, F., Shvaiko, P., and Yatskevich, M.: Semantic Schema Matching. 13th International Conference on Cooperative Information Systems (CoopIS 2005).
9. Hearst, M., English, J., Sinha, R., Swearingen, K. and Yee, P.: Finding the Flow in Web Site Search. Communications of the ACM, 45 (9), 2002.
10. Hyvönen, E., Mäkelä, E., Salminen, M., Valo, A. et al.: MuseumFinland - Finnish Museums on the Semantic Web. Journal of Web Semantics, 3(2), 2005.
11. Jian, N., Hu, W., Cheng, G., and Qu, Y.: Falcon-AO: Aligning Ontologies with Falcon. K-CAP Workshop on Integrating Ontologies, Banff, Canada, 2005.
12. Kalfoglou, Y., Schorlemmer, M.: Ontology Mapping: The State of the Art. The Knowledge Engineering Review Journal, 18(1), 2003.

A Content-Based Image Retrieval Service for Archaeology Collections

Naga Srinivas Vemuri[1], Ricardo da S. Torres[2], Rao Shen[1],
Marcos André Gonçalves[3], Weiguo Fan[1],
and Edward A. Fox[1]

[1] Digital Library Research Lab, Virginia Tech, USA
{nvemuri, rshen, mgoncalv, wfan, fox} @vt.edu
[2] Institute of Computing, State University of Campinas, Av. Albert Einstein, 1251,
CEP 13084-851, Campinas, SP, Brazil
rtorres@ic.unicamp.br
[3] Department of Computer Science, Federal University of Minas Gerais,
CEP 31270-901, Belo Horizonte, MG, Brazil
mgoncalv@dcc.ufmg.br

Abstract. Archeological sites have heterogeneous information ranging from different artifacts, image data, geo-spatial information, chronological data, and other relevant metadata. ETANA-DL, an archaeology digital library, provides various services by integrating the heterogeneous data available in different collections. This demonstration presents an initial prototype for searching DL objects based on the image content, using the Content-Based Image Search Component (CBISC) from Virginia Tech/State University of Campinas.

1 Introduction

Archeological systems involve heterogeneous data such as different kinds of artifacts, corresponding images, geo-spatial information, chronological information, and relevant metadata. ETANA-DL [1], an archaeology digital library, tries to integrate this heterogeneous metadata and provides services to its user societies. Archeologists consider artifact's image data as vital information for documenting, analyzing, and sharing. We address this fact by developing an initial search prototype that uses a Content-Based Image Search Component (CBISC) recently proposed [2].

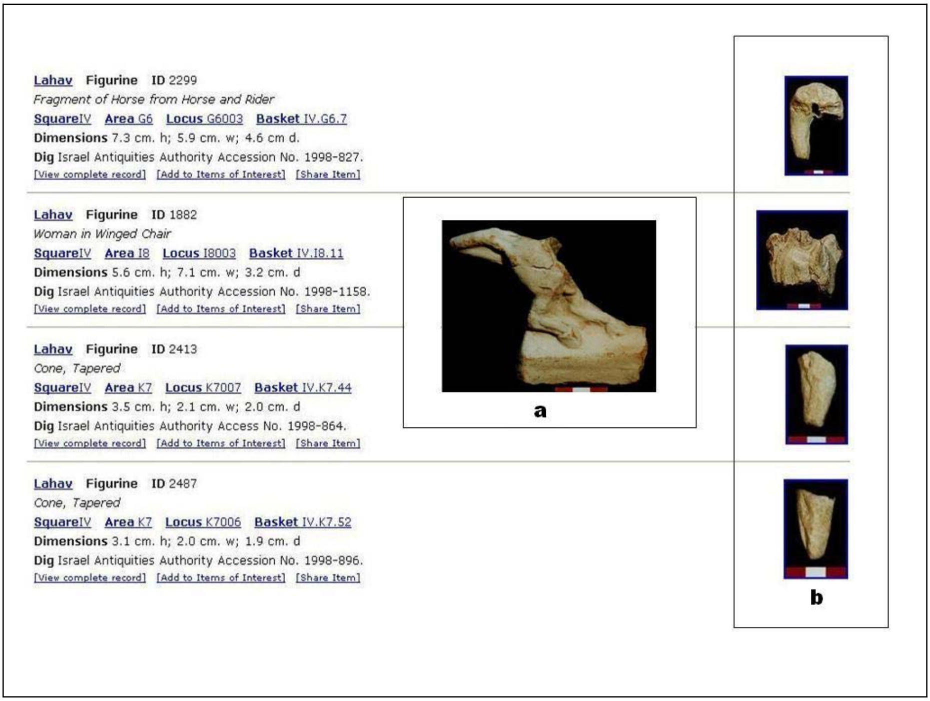
2 Our Approach

The CBISC prototype allows the end user to search for similar DL objects based on the image content present in the system. To perform a query, the user has to upload the query image and specify k to denote the desired number of similar images. The CBISC component extracts the feature vectors from the query image using the Border/Interior Pixel Classification [3] image descriptor, computes L_1 distance with feature vectors of the ETANA-DL image collection, and returns the top k DL artifacts whose image content is similar to that of the query. Figure 1a shows a sample query

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 438 – 440, 2006.

image, and Figure 1b shows the top 4 similar images returned for the query image. We identify that these returned images appear more relevant compared to the other images present in the collection. The entire figure shows the corresponding DL objects returned by the CBISC component. We used the BIC image descriptor, as it suits well with the kind of images available in our system.

Our approach is different from other content based retrieval approaches in that the complete DL object information along with its image content is returned for a query image. We believe that this approach will be very useful especially when collections have inconsistencies in their metadata. From our past experience, it is not uncommon for archaeological collections to have inconsistencies in textual metadata. In such a situation, an archaeologist looking at an artifact in a particular dig might discover other artifacts discovered at the same dig even though the corresponding textual metadata is incorrect. It should be noted that this is more a complimentary strategy than an alternative to textual metadata search.



Lahav **Figurine** **ID** 2299
Fragment of Horse from Horse and Rider
SquareIV **Area** G6 **Locus** G6003 **Basket** IV.G6.7
Dimensions 7.3 cm. h; 5.9 cm. w; 4.6 cm d.
Dig Israel Antiquities Authority Accession No. 1998-827.
[View complete record] [Add to Items of Interest] [Share Item]

Lahav **Figurine** **ID** 1882
Woman in Winged Chair
SquareIV **Area** I8 **Locus** I8003 **Basket** IV.I8.11
Dimensions 5.6 cm. h; 7.1 cm. w; 3.2 cm. d
Dig Israel Antiquities Authority Accession No. 1998-1158.
[View complete record] [Add to Items of Interest] [Share Item]

Lahav **Figurine** **ID** 2413
Cone, Tapered
SquareIV **Area** K7 **Locus** K7007 **Basket** IV.K7.44
Dimensions 3.5 cm. h; 2.1 cm. w; 2.0 cm. d
Dig Israel Antiquities Authority Access No. 1998-864.
[View complete record] [Add to Items of Interest] [Share Item]

Lahav **Figurine** **ID** 2487
Cone, Tapered
SquareIV **Area** K7 **Locus** K7006 **Basket** IV.K7.52
Dimensions 3.1 cm. h; 2.0 cm. w; 1.9 cm. d
Dig Israel Antiquities Authority Accession No. 1998-896.
[View complete record] [Add to Items of Interest] [Share Item]

Fig. 1. An example scenario of content-based image query in ETANA-DL

3 Conclusion

Presently, our focus is to extend this architecture, before the conference, by providing other services on top of this component. The existing recommendation component of ETANA-DL provides recommendations to individual users using a collaborative filtering mechanism. We shall extend it to provide recommendations based on the image content of a selected DL object. Also, we shall perform a comparison with other image descriptors to evaluate their effectiveness for archaeology image collections. After deployment, we shall evaluate these services by performing usability studies.

Acknowledgements. This work is funded in part by the National Science Foundation (ITR-0325579). Further support is provided through CNPq, CAPES, FAPESP, CNPq WebMaps and AgroFlow projects, FAEPEX, 5S-QV project grant MCT/CNPq/ CT-INFO 551013/2005-2 and by the Microsoft eScience grant.

References

1. U. Ravindranathan. Prototyping Digital Libraries Handling Heterogeneous Data Sources - An ETANA-DL Case Study. Masters Thesis. Computer Science, Virginia Tech, Blacksburg VA, April 2004, http://scholar.lib.vt.edu/theses/available/etd-04262004-153555/
2. Torres, R. da S., Medeiros, C.B., Goncalves, M.A., Fox, E.A.: A digital library framework for biodiversity information systems. International Journal on Digital Libraries 6 (2006) 3-17
3. Stehling, R. O., Nascimento, M. A., Falcão, A. X.: A compact and efficient image retrieval approach based on border/interior pixel classification. CIKM 2002: 102-109.

A Hierarchical Query Clustering Algorithm for Collaborative Querying

Lin Fu, Dion Hoe-Lian Goh, and Schubert Shou-Boon Foo

Division of Information Studies
School of Communication and Information
Nanyang Technological University
Singapore 637718
{fulin, ashlgoh, assfoo}@ntu.edu.sg

Abstract. In this work, a hierarchical query clustering algorithm is designed and evaluated for the collaborative querying environment. The evaluation focuses on domain specific queries to better understand whether the algorithm meets the needs of users. Experiment results show that the proposed algorithm works well to cluster queries with good precision.

1 Introduction

Collaborative querying addresses the issue of query formulation by sharing other users' search experiences to help users formulate appropriate queries to a search engine. In our previous work [3], a collaborative querying system was developed to assist in query formulation by finding previously submitted similar queries through mining query logs. The system operates by clustering and recommending related queries to users. Since similarity is fundamental to the definition of a cluster, measures of similarity between two queries are essential to the query clustering procedure. We propose a hybrid query similarity measure that exploits both the query terms and query results URLs. Experiments reveal that using the hybrid approach, more balanced query clusters can be generated than using other techniques [2].

Besides similarity measures, the clustering algorithm is another key factor that affects the quality of the query clusters. In this paper, we describe a hierarchical query clustering algorithm based on the hybrid similarity measure. Further we report the experiments results of the proposed algorithm on domain specific queries.

The rest of this paper is organized as follows. In Section 2, we present the design and implementation of the hierarchical query clustering algorithm as well as the experiments for the algorithm. Then, we discuss the implications of our findings for collaborative querying.

2 A Hierarchical Query Clustering Algorithm

Given a query, *Qi,* submitted by a user, our approach is to detect the query cluster G(*Qi*) containing the initial query *Qi* as shown in Table 1.

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 441 – 444, 2006.

Table 1. Direct query cluster detection algorithm

```
1. Direct (Qi , N, T)
2. If the query repository contains the given query, Qi
3.      For each query, q', in the query repository that is related to Qi,
4.      If sim_hybrid(Qi,q')>=T,
5.            Add q' to List G(Qi)
6.      Rank the members according to their similarity to the query Qi and delete
   those out of the top N members.
7. Return G(Qi)
```

The inputs of the algorithm are *Qi* , *N* and *T* as shown in the first line. *Qi* is the query submitted by the user. *N* denotes the maximum size of the query cluster and *T* is the threshold for query clustering. After obtaining the query cluster G(*Qi*), the system will recommend the queries in the cluster to users. Once the user selects one of the recommended queries by clicking on it, the system will repeat the algorithm. Here the G(*Qi*) returned contains the queries are that directly related to *Qi*.

If the user wants to activate the QGV to visualize the related queries to the submitted query *Qi*, the system will further detect query clusters containing the members of G(*Qi*) according to the following algorithm shown in Table 2.

Table 2. Indirect query clusters detection algorithm

```
1. Indirect (G(Qi), M)
2. For each query, q', in G(Qi)
3.        Direct (q', N, T)
4.        Add G(q') to G'(Qi)
5.        M=M-1
6.        if M>=0, Indirect (G(q'), M)
7. return G'(Qi)
```

The inputs to the algorithm are G(*Qi*) and *M* as shown in the first line. *Qi* is the query submitted by the user. *M* denotes the maximum level of query clusters to be retrieved from the *Qi*. Here the cluster G'(Qi) contains queries that are indirectly related to the query *Qi*.

Consider for example the query "data mining". Suppose this query is matched to the cluster {"data mining", "predictive data mining", "knowledge discovery", "data mining applications", …}. This set of queries will form the level 1 nodes around the root "data mining". The algorithm will then expand all the individual members of this initial cluster to retrieve indirectly related queries. Suppose "knowledge discovery" is currently being expanded and results in the cluster {"data warehousing, OLAP, data mining", "data mining journal", …}. These queries will form the level 2 nodes around the "knowledge discovery" node. Figure 1 presents parts of the hierarchically organized query clusters based on the root "data mining".

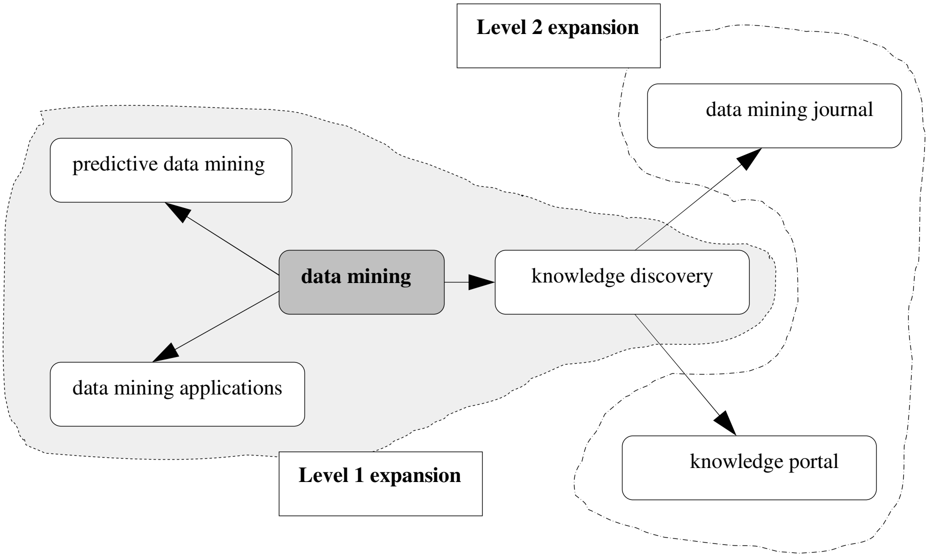


Fig.1. Parts of related query clusters for "data mining"

3 Evaluation and Results

Eight participants were divided into four groups of two participants each according to their expertise. The two participants in each group were asked to evaluate the precision of the query clusters generated for one domain. Here the precision of query clusters refers to the ratio of the number of similar queries to the total number of queries in a cluster.

The evaluation relied entirely on subjective judgments by the two domain experts. Therefore, an assessment of the evaluators' degree of agreement in evaluation would indicate the reliability of their judgment. Cohen's Kappa statistic was adopted to measure the degree of agreement between the two evaluators' judgments [1]. Kappa has a range from 0 to 1, with 1 indicating perfect agreement and 0 indicating poorly agreement. Usually, a Kappa value of more than 0.8 is considered satisfactory [4]. Table 3 shows the results of this study. The average precision was 65.62%, 66.83%, 68.76% and 62.69% respectively for the domains of data mining, knowledge management, cross cultural communication and wireless communication. This result shows that the majority queries in the query clusters are relevant to the original query. Note that the precision can be boosted by using higher threshold values to cluster queries. The average precision across these four domains was 65.98% which was higher than the previous precision value (49.28%) reported in [2]. The reason behind this is that the evaluators in this study were domain experts and were familiar with domain specific query terms. As such, when there were abbreviations (for example, OLAP) or human names (for example, han jiawei), the evaluators could provide more accurate judgments on the relevancy between queries. The Kappa statistic was 0.81, 0.95, 0.82 and 0.86 respectively for the domains of data mining, knowledge management, cross cultural communication and wireless communication. This result shows a high degree of agreement between the evaluators for each domain. In other words, the results obtained from the evaluation were reliable.

Table 3. Average precision and Kappa statistic for domain specific query clusters

Domain	Average Precision	Average Kappa
Data Mining	65.62%	0.81 ($p < 0.001$)
Knowledge Management	66.83%	0.95 ($p < 0.001$)
Cross cultural Communication	68.76%	0.82 ($p < 0.001$)
Wireless Communication	62.69%	0.86 ($p < 0.001$)

References

[1] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46.

[2] Fu, L. Goh, D. & Foo, S. (2003). Collaborative querying through a hybrid query clustering approach. *Proceedings of Sixth International Conference of Asian Digital Libraries*, 111-122.

[3] Fu, L. Goh, D. & Foo, S. (2004). Collaborative querying for enhanced information retrieval. *Proceedings of ECDL2004*, 111-122.

[4] Krippendorff, K. (1980). *Content Analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, CA: Sage Publications.

A Semantics-Based Graph for the Bib-1 Access Points of the Z39.50 Protocol*

Michalis Sfakakis and Sarantos Kapidakis

Archive and Library Sciences Department / Ionian University
Plateia Eleftherias, Paleo Anaktoro GR-49100 Corfu, Greece
{sfakakis, sarantos}@ionio.gr

Abstract. A graph of Access Points can be used as a tool in a number of applications such as, clarification and better understanding of their semantics and inter-relations, query transformations, more precise query formulation, etc. We apply a procedure to create a graph of the Access Points, according to their subset relationship, based on the official semantics of the Bib-1 Access Points of the Z39.50 protocol. In our three-step method, we first construct the relationship graph of the Access Points by testing for subset relationship between any two Access Points, and assigning each Access Point a weight value designating the number of the Access Points, which are subsets to it. In the second step, we apply a topological sorting algorithm on the graph, and finally in the last step, we reject the redundant subset relationships of the Access Points.

1 Introduction

The query mechanism of the Z39.50 [1] protocol, utilizes sets of predefined Access Points combined with specific attributes (i.e. Attribute Sets), in a number of different query language specifications (i.e. query types). One of the elements defined in an Attribute Set is the set of the valid Access Points (i.e. what entities represent the search terms, like Title, Author, etc.) from a specific set of attribute types. The Bib-1 Attribute Set is the most commonly used one, and provides the *Use* attribute type for the specification of an Access Point against which the search term is to be matched.

A semantics-based Access Point graph is necessary to better understand the exact semantics of every Access Point, as well as their inter-relationships, and can be used in a number of cases: When query optimization is involved, the use of the graph could help the transformation process of the query. Also, it could be a helpful tool for automated, or semi automated, procedures when either Access Point replacement is required, due to unsupported Access Points in a query, or Z39.50 queries are used in a heterogeneous information retrieval environment. For the end user, a better understanding of the quality of the search terms that arises from such a graph, could guide him to pose more precise (interactive) queries.

* This work was supported by the DELOS Network of Excellence on Digital Libraries (EU 6. FP IST, G038-507618).

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 445 – 448, 2006.

We generate the graph of the Access Points according to the subset relationship among them, based on the MARC elements used on the definition for the semantics of the Bib-1 Access Points [2]. We recall that, the semantics of the Bib-1 Access Points implement the search requirements posed by the user community of the Z39.50 protocol. Also, the existence of the XML syntax (MARCXML) of the MARC 21 specification does not affect the derived graph, due to unchanged semantics.

Highlighting some of our results, shown in fig. 1, the Access Point *Any* is the superset of all the other Access Points. The largest component of the graph with the longest path starts from the *Author-Title-Subject*, which has as subset the *Name*, which has as subset the *Author-name*, which has as subset the *Author-name-personal*, which finally has as subset the *Name-editor* and is also linked to *Name-personal* to which is a subset.

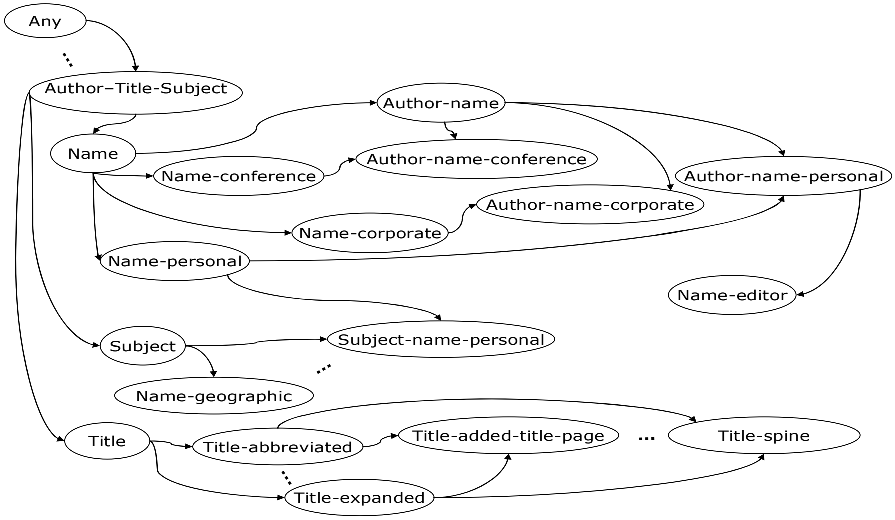


Fig. 1. A representative sample of the semantics-based Bib-1 Access Points graph

2 Method Description

The procedure to create the graph for the interrelations of the Access Points consists of three steps. Initially in the first step, we create the relationship graph of the Access Points by testing for subset relationship between all pairs of Access Points, and assigning to each Access Point a weight value designating the number of the Access Points which are subsets to it. In the second step, we apply a topological sorting algorithm on the graph. Finally, in the third step, we reject the redundant subset relationships by keeping the longest path between every pair of connected Access Points. We consider an explicit relationship as redundant, if we are able to infer its existence from other relationships of the Access Points.

We consider an Access Point to be a subset of another, if the set of the data fields used to create the first is a subset of the set of the data fields used to create the second.

As an example, consider the Access Point *Author-name* which, according to its definition [2], includes the data from the fields with MARC tags included in the set {100, 110, 111, 400, 410, 411, 700, 710, 711, 800, 810, 811}, and also, the Access Point *Author-name-personal* which includes the data from the set of fields {100, 400, 700, 800}. The Access Point *Author-name- personal* is considered being a subset of the *Author-name*.

We represent the relationships between the Access Points with a directed graph G in which the vertices represent Access Points and the edges represent subset relationships. This graph has an edge <*i, j*> if and only if Access Point *i* is a subset of Access Point *j*. The Access Points *Author-name* and *Author-name-personal*, used in the previous example, will be represented by two vertices of the graph and their subset relationship from the edge <*Author-name-personal, Author-name*>. The out-degree of a vertex expresses the number of the subsets for the represented Access Point by the vertex, as specified by the semantics definition of the Bib-1 Access Points.

The following example will better clarify our method. Let's consider that the Bib-1 Attribute Set consists only of the next four Access Points: The *Any*, the *Abstract*, the *Data-acquisition* and the *Note* Access Point. According to the Bib-1 semantics specification, the *Any* Access Point can be thought as the union of all the supported Access Points (i.e. a superset of all the others). The *Abstract* Access Point includes the data from the set of field {520}, the *Data-acquisition* includes the data from the set of field {541-subfield-d}, and finally, the *Note* Access Point includes the data from the set of fields {500, 501, …, 520, …, 535, 536, …, 541, …, 586}. We can see that all the Access Points are subsets to *Any*, and also that, the Access Points *Abstract* and *Data-acquisition* are subsets of the *Note* Access Point. Using these interrelations of the Access Points, we construct the graph G shown in fig. 2, thus completing the first step of our method.

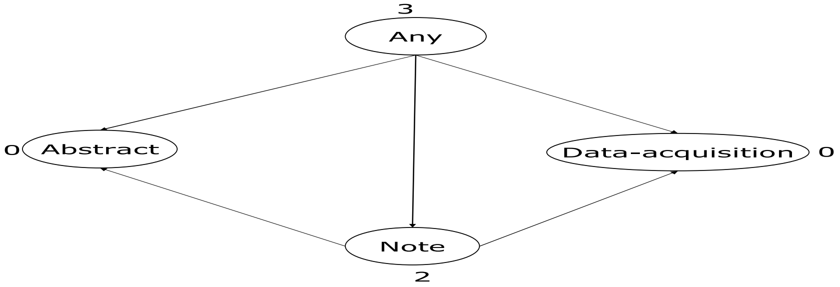


Fig. 2. Step 1: Construction of the G graph. The number near a vertex expresses its out-degree.

After applying the topological sorting algorithm on the graph, step 2, we rearrange the graph as shown in fig. 3. Obviously, this ordering is feasible due to the transitive and irreflexive properties of the proper subset relation.

At the last step we delete the derivative subset relationships. We number the vertices from left to right and for each vertex, we only keep the incoming edge from the highest numbered vertex. The resulting Graph, G0, is the minimal subset of the initial graph G, so that the transitive closure of G0 produces the graph G. The final arrangement of the graph is shown in fig. 4.

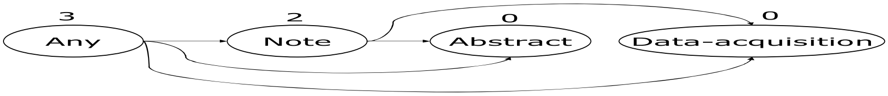


Fig. 3. Step 2: Graph G after the topological sorting

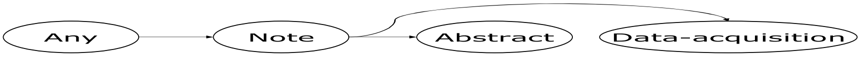


Fig. 4. Step 3: Graph G0, the minimal subset of G

3 Discussion

An important practical use of our results is when Access Point replacement is required due to unsupported Access Points in a Z39.50 query. This case is very common when we query many different Z39.50 servers. The following example illustrates some real world circumstances when a client tries to accomplish a parallel search in many sources, and also how the client could use the Access Point graph. Consider two sources, where the first one supports the Access Point *Author-name* and the second one supports the Access Point *Author-name-personal.* Obviously, all requests to the first server for selecting data using the Access Point *Author-name-personal* will fail. A smart client could substitute the Access Point *Author-name-personal* with the Access Point *Author-name* into the queries, taking into account that the Access Point *Author-name-personal* is a subset of the Access Point *Author-name*. This way, the client could avoid the failure of the query, although, unavoidably, the precision of the resulting query will be less than the precision of the original one. In this example we made the assumption that both sources support the same value combinations for the remaining attribute types (i.e. *Relation*, *Position*, *Completeness, etc.*), in order to simplify its description.

References

1. ANSI/NISO: Z39.50 Information Retrieval: application service definition and protocol specification: approved May 10, 1995.
2. Attribute Set BIB-1 (Z39.50-1995): Semantics. ftp://ftp.loc.gov/pub/z3950/defs/bib1.txt.

A Sociotechnical Framework for Evaluating a Large-Scale Distributed Educational Digital Library

Michael Khoo

National Science Digital Library, P.O. Box 3000, Boulder,
CO 80307-3000, U.S.A.
mjkhoo@ucar.edu

Abstract. The National Science Digital Library (NSDL: http://www.nsdl.org) supports all levels of science, technology, engineering, and mathematics education. NSDL is conducting a program-wide evaluation of all its activities since 2000. The scale and complexity of the NSDL program pose significant challenges for this evaluation work. This poster outlines a sociotechnical theoretical framework, the 'resource lifecycle,' that is being used to guide the evaluation of the NSDL program.

1 The National Science Digital Library (NSDL)

The National Science Digital Library (NSDL: http://www.nsdl.org) is a United States National Science Foundation program that is building distributed digital library infrastructure to support the teaching of science, technology, engineering and mathematics (STEM). NSDL supports all educational levels, from kindergarten through high school and university to lifelong learning, and has funded over 200 projects since 2000. NSDL is conducting a program-wide evaluation, with the aim of identifying important strategic areas for future library development.

This evaluation faces a number of challenges. First, the program is widely distributed, with projects located across the United States; and while several 'Core Integration' projects are tasked with coordinating the development of NSDL there is no program center to which all projects are required to report. Second, many projects are funded by NSF on a one-shot basis; and since the average project funding time is 2-3 years, and NSDL made its first funding awards in 2000, this means that many projects have been completed and their staff have moved onto other work. Third, NSDL has funded a wide range of research ranging from database and search engine design, to resource creation and metadata generation and collection development, to website design, and to community and outreach activities; there is thus no typical NSDL project, and no standard evaluation metrics and methodologies applicable across the entire program. Fourth, NSDL projects include a range of 'communities of practice,' including computer scientists, librarians, and STEM teachers, who can have different and sometimes competing definitions of digital libraries [2]; and the evaluation work will have to address the concerns of these different groups. Finally, resources for NSDL evaluation are limited, both financially, and in terms of time.

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 449 – 452, 2006.

Successful digital library evaluation depends on the identification of ‘doable’ evaluation questions, the outcomes of which can be used to guide a digital library’s long-term development [5]. Evaluation approaches should be *multifaceted*, embodying multiple data views that capture the complexity of digital library technologies and user behaviours [3]. At the same time, as has been described, the NSDL program has a a limited evaluation budget, within which it is impossible to examine every dimension of every NSDL project. Which dimensions, therefore, should be evaluated? A useful approach here is to consider NSDL as a sociotechnical system, that is, as a complex system of technologies, people, practices, and other phenomena, that are related in emergent and unpredictable ways [1, 6]. A sociotechnical approach models these phenomena within a systematic framework, and provides accounts of how these components may be interrelated. The approach works well when the researcher focuses on one phenomenon as the unit of analysis and ‘follows’ this unit throughout the sociotechnical system, using the unit to construct of models of local areas of the system in question.

The sociotechnical model developed to guide NSDL program evaluation is called the ‘resource lifecycle’ model [4]. The model assumes that the overall purpose of NSDL is to transform scientific data into useful pedagogical resources, through a coordinated series of value-adding activities. In following and describing this process of transformation, the chosen unit of analysis is that of the educational resource. In the model, educational resources exist in a number of different forms, and in a range of different contexts, and are subject to different actions by different groups, aimed at improving the utility and value of these resources. Each of these value-adding activities involves a range of local actions, and the identification of these actions allows in turn the specification and development of multifaceted [3] evaluation frameworks and metrics to measure the efficacy of these actions. One example of a value-adding activity is that of educational resource creation. Creating high quality resources is identified as a central task for NSDL, and relevant evaluation questions here thus include: What resource creation guidelines did individual projects follow? What quality assurance processes did they follow? And, how did they rate the support that they received from NSDL in carrying out these activities?

The ‘resource lifecycle’ model identifies five core interlinked NSDL value-adding activities (Figure 1): (a) resource creation and review; (b) resource aggregation and collection development; (c) web site and search engine design; (d) use and reuse; and (e) organizational communication and knowledge (meetings, e-mail lists, wikis, etc.). These five basic areas provide a framework for the identification of more specific areas of NSDL activity, such as: the creation and review of educational resources; resource aggregation and the generation of item-level and collection-level metadata; the design of the NSDL portal and usability testing; resource search and discovery on nsdl.org; outreach and publicity efforts; support for NSDL users to share and re-use resources; and so on. Each of these stages involves, from a sociotechnical point of view, a mixture of technologies, people and practices that are amenable to further analysis, investigation and evaluation. Identifying these stages permits the evaluation of NSDL not on a project-by-project but rather a cross-cutting basis that can include many NSDL projects in a single evaluation activity. Rather than evaluating each

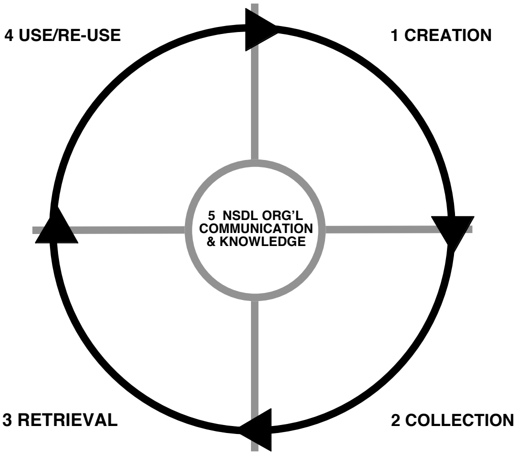


Fig. 1. The Resource Lifecycle (overview)

individual project, the model looks to how the 200 projects within NSDL collaborate on common tasks in order to fulfill the overall mission of NSDL.

1.1 Example NSDL Evaluation Activities

Evaluation activities have been initiated in all the NSDL core activity areas identified by the model. A summary of these activities follows; full reports of these activities are on the NSDL evaluation wiki: http://eval.comm.nsdl.org/.

Resource creation and review (core activity 1) Resource creation and review processes in NSDL projects were the subject of a web survey e-mailed to all past and present NSDL projects. The highest response rates came from the most recently funded projects, confirming the problem of reaching older NSDL projects. Responses suggested that the majority of projects did follow resource creation guidelines, and some projects supplied copies of their guidelines, to be made available to other projects. An unanticipated outcome was that many respondents ranked NSDL's organizational communication infrastructure (e-mail lists, wikis, etc.) as of low importance; and steps are being taken to improve these services.

Metadata (core activity 2) Metadata quality is an important component of the resource lifecycle model, with better quality metadata supporting better search results for users. Initial analyses are being conducted with the NSDL Metadata Repository (MR), and the soon-to-be-launched NSDL Data Repository (NDR), to assess such factors as field completion rates, and the quality of data in specific fields. The results of these analyses will be used to inform search engine and search page design.

Search page usability (core activity 3) In conjunction with metadata quality analysis, user-testing of the search functions of the nsdl.org search pages is being carried out. Research is currently in the paper prototyping stage, the results of which will inform more comprehensive user-testing by an external HCI specialist later in 2006.

Webmetrics (core activity 4) Individual NSDL projects often carry out their own web metrics, with little standardization in tools or results. NSDL has contracted standardized cross-project third-party web metrics from Omniture (omniture.com), in order to identify typical patterns of use across NSDL, and to inform future web site design (for instance by identifying heavily- and lightly-used sectors of the NSDL web site).

Project-level evaluation practices (core activities 1-4) An online survey designed to collect individual projects' own evaluation activities indicated that while projects are eager to carry out this work, they often face obstacles based on lack of funds, time, and expertise. This suggests the need for some kind of centralized evaluation service that could be provided by NSDL, which could provide contract evaluation services for individual projects.

Annual Meeting and workshops (core activity 5) NSDL activities such as the NSDL Annual Meeting and workshops help to support community amongst NSDL projects, and provide opportunities for new collaborations and innovation. Paper and online surveys conducted after the 2005 Annual Meeting suggested that NSDL projects strongly value opportunities to meet and interact face-to-face, and to develop their knowledge and possibilities for collaboration. These findings replicate the findings of other NSDL surveys which similarly have commented on the importance to individual projects of the overall NSDL community.

2 Summary

The resource lifecycle model is being used by NSDL to inform the design of a multi-faceted range of evaluation activities. As a model, it appeals to and makes sense across a range of NSDL communities of practice, and provides useful conceptual boundaries within which evaluation efforts may be developed and applied.

References

1. Bijker, W. 1995. Of bicycles, bakelites and bulbs. Toward a theory of sociotechnical change. Cambridge, MA: The MIT Press.
2. Khoo, M. 2005. The tacit dimensions of user behavior: The case of the Digital Water Education Library. JCDL 2005, pp. 213-222. New York: ACM Press.
3. Marchionini, G., C. Plaisant, and A. Komlodi. 2003. The people in digital libraries: Multifaceted approaches to assessing needs and impact. In: Bishop, A., N. van House, and B. Buttenfield (eds.), *Digital Library Use: Social Practice in Design and Evaluation.* Cambridge, MA: The MIT Press, pp. 119-160.
4. NSDL. 2006. The 'Resource Lifecycle': A conceptual metaframework for NSDL program evaluation. http://eval.comm.nsdl.org/cgi-bin/wiki.pl?WhitePaper
5. Reeves, T., X. Apedoe, & Y. H. Woo. 2003. Evaluating Digital Libraries: A User-Friendly Guide. UCAR. http://dlist.sir.arizona.edu/398/
6. Van House, N., A. Bishop, and B. Buttenfield. 2003. Digital libraries as sociotechnical systems. In: Bishop, A., N. van House, and B. Buttenfield (eds.), *Digital Library Use: Social Practice in Design and Evaluation.* Cambridge, MA: The MIT Press, pp. 1-21.

A Tool for Converting from MARC to FRBR

Trond Aalberg[1], Frank Berg Haugen[2], and Ole Husby[3]

[1] The Norwegian University of Science and Technology (NTNU), IDI
[2] The National Library of Norway
[3] BIBSYS, Norway

Abstract. The FRBR model is by many considered to be an important contribution to the next generation of bibliographic catalogues, but a major challenge for the library community is how to use this model on already existing MARC-based bibliographic catalogues. This problem requires a solution for the interpretation and conversion of MARC records, and a tool for this kind of conversion is developed as a part of the Norwegian BIBSYS FRBR project. The tool is based on a systematic approach to the interpretation and conversion process and is designed to be adaptable to the rules applied in different catalogues.

1 Motivation

The Entity Relationship (ER) model defined in the *Functional Requirements for Bibliographic Records* [7] aims to enable libraries to meet the broad range of increasing user expectations and needs, and the model is considered by many to be a major contribution to the next generation of bibliographic catalogues. The core of the model is the use of the entities *work*, *expression*, *manifestation* and *item* to model the products of intellectual and artistic endeavour at different levels of abstraction.

Due to the very large number of already existing bibliographic records in the MARC-format, a major challenge for the library community is to be able to apply the FRBR model on existing bibliographic catalogues. This is commonly referred to as "frbrization" and requires a solution for the automatic interpretation and conversion of existing information. A tool for this purpose has been developed in a joint frbrization project between The Norwegian University of Science and Technology, The Norwegian bibliographic database BIBSYS and the National Library of Norway[1]. The tool supports the full conversion of a MARC-based bibliographic catalogue into a format that directly reflects the FRBR model. A systematic approach to the conversion process is used in the design of the tool and the conditions that govern the conversion process are stored externally. This approach facilitate easier definition and consistent maintenance of the conversion and additionally enables reuse of the tool across catalogues that may be different in terms of cataloguing practise and the use of MARC format.

[1] The project was funded by the Norwegian Archive, Library and Museum Authority.

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 453–456, 2006.

2 Interpreting and Processing MARC Records

At the most generic level the process of interpreting a MARC record in the context of the FRBR model consists of (1) identifying the various entities described in the record, (2) selecting the MARC fields that describe each entity and (3) finding the correct relationships between the entities. Each FRBR entity may appear in a bibliographic record with different roles; e.g. persons can be found as the creator of a work or as a subject of the work, and with many other roles as well. To account for all the combinations of entities and their roles, a conversion process needs to be able to deal with a rather large number of cases. The need for selecting and assigning MARC fields to the correct entities and establishing the proper relationships between entities further complicates this problem.

3 The Conversion Tool

The process outlined above is implemented in our conversion tool by the use of XSLT – the W3C language for transforming XML. Each entity case is coded as a template following the same control structure, and XPATH expressions are used for the various conditions and selections that need to be applied. The tool reads MARC-records encoded in the MarcXchange XML-format [4] and produces a record for each entity in a format that extends the MarcXchange with FRBR type attributes and a relationship element.

The basic structure of each template consists of an outer for-each loop with an entity condition that tests for the existence of data fields or other information that indicates the presence of an entity (or multiple entities). The code that creates the record is inside this loop and includes the generation of an identifying key, the code that selects and copies the appropriate data fields from the source record, and the code that creates the proper relationships to other entities.

The tool is further generalized by the use of a database for storing the data that is variable for each template. This includes the entity conditions, the mapping of entities from MARC to FRBR and the conditions and targets for relationships. The various templates needed in the conversion are created by a program written for this purpose. When the templates have been created the conversion can be run using code that iterates over a collection of records and applies all entity templates on each record. This solution enables reuse of the tool even across catalogues that have different use of the MARC format. Rather than having to implement a conversion tool ad-hoc, the same tool can be used but with different rules.

The process outlined so far is only concerned with the interpretation and conversion of each record into a corresponding set of interrelated FRBR entities. In bibliographic catalogues the same entity is often duplicated across a number of records, and a merging of identical entities is required to create a normalized frbrization with a consistent set of relationships. The merging process is performed after the actual conversion and is based on the use of descriptive entity keys. Each entity is identified by keys that are created from a set of identifying field values. Equal keys will be generated for entities that are described by the same set of identifying field values, but a tolerant key matching algorithms can be applied to allow for certain variations in the data.

In addition to the core process of creating FRBR entities and relationships, a complete conversion typically needs to include both a preprocessing and postprocessing step. Preprocessing is often needed to enable different kinds of preparations that are more conveniently solved in advance rather than during the conversion. An additional postprocessing step is included as well to support the possible processing of the final results into other formats or to enrich the result with additional data. Various preprocessing and postprocessing tasks can easily be added to the conversion tool by including additional XSLT templates that processes either the input format before the frbrization or the output from the conversion.

4 Status, Related Work and Future Issues

The conversion tool has successfully been used to convert the 4 million records in the BIBSYS bibliographic database. The result has been used in a prototype catalogue available on the web. This project has verified that the conversion tool is able to capture and apply all the different conditions and rules that govern the interpretation of the BIBSYS-MARC format.

The tool can be used in a variety of settings. Currently there are many libraries that want to do frbrization as an experimental project to learn about the FRBR model or to test how well the records can be converted, and the tool may serve this need. A full conversion to a FRBR-based data model is probably not realistic for most libraries yet due to the limited knowledge about the actual usability of the model and the lack of standardized formats for FRBR, but the conversion tool can be used in real-world applications in other ways such as for creating indexes that can be used to support frbrized views on existing catalogues.

The main problems uncovered in the BIBSYS frbrization project are the efficiency of a large scale conversion based on XSLT and the problem of dealing with inconsistent data. The results from our conversion demonstrate that a perfect set of FRBR entities and relationships can be produced if the initial records contain sufficient and consistent information. For other sets of records the conversion tool creates duplicate entities and erroneous relationships, but these problems can usually be ascribed to inconsistencies and errors in the initial records rather than the conversion tool.

The conversion tool is influenced by comparable work by others such as the identification of FRBR entities in bibliographic catalogues which have been explored in [1,2,3,5]. Our mapping of MARC fields to FRBR entities and attributes follows the same pattern as the mapping that is defined for the MARC 21 format in [9], but additionally includes conditions which enables a more detailed definition. A few other tools for experimenting with frbrization already exist such as the FRBR display tool made available by The Library of Congress Network Development and MARC Standards Office [8], and the workset algorithm developed by OCLC [6]. The key algorithm applied in this tool is comparable to the workset algorithm of OCLC but is applied on all entities. Compared to the work of others the main contribution of our tool is the systematic approach to the conversion process and the subsequent support for creating a complete frbrization of MARC records.

The tool is being further developed at NTNU, and is planned to be made freely available. The use of the tool to do large-scale conversions of other MARC-formats is needed to verify the reusability of the system. Further work on the tool includes the implementation of support utilities and services than can be used to solve the more difficult problems of identifying comparable entities e.g. by the use of external authority files and specific purpose extension functions for comparing entity keys. Additional future work includes the conversion to other formats; particularly the use of RDF combined with a formal FRBR ontology e.g. in OWL will be explored.

References

1. Marie-Louise Ayres. Case studies in implementing Functional Requirements for Bibliographic Records [FRBR]: AustLit and MusicAustralia. *ALJ: the Australian Library Journal*, 54(1):43–54, February 2005. http://www.nla.gov.au/nla/staffpaper/2005/ayres1.html.
2. Knut Hegna and Eeva Murtomaa. *Data Mining MARC to Find : FRBR?* BIBSYS/HUL, 2002. http://folk.uio.no/knuthe/dok/frbr/datamining.pdf.
3. Thomas B. Hickey, Edward T. O'Neill, and Jenny Toves. Experiments with the IFLA Functional Requirements for Bibliographic Records (FRBR). *D-Lib Magazine*, 8(9), September 2002. http://www.dlib.org/dlib/september02/hickey/09hickey.html.
4. International Organization for Standardization TC46/SC4. Information and Documentation : MarcXchange. Draft standard ISO/CD 25577, ISO, 2005. http://www.bs.dk/marcxchange/.
5. Christian Mönch and Trond Aalberg. Automatic conversion from MARC to FRBR. In *Research and Advanced Technology for Digital Libraries, ECDL 2003*, number 2769 in Lecture Notes in Computer Science, pages 405–411. Springer-Verlag, 2003.
6. OCLC. FRBR work-set algorithm. http://www.oclc.org/research/projects/frbr/algorithm.htm.
7. IFLA Study Group on the functional requirements for bibliographic records. *Functional requirements for bibliographic records : final report*, volume 19 of *UBCIM Publications : New Series.* K. G. Saur, Munich, 1998. http://www.ifla.org/VII/s13/frbr/frbr.pdf.
8. The Library of Congress' Network Development and MARC Standards Office. FRBR Display Tool. http://www.loc.gov/marc/marc-functional-analysis/tool.html.
9. The Library of Congress' Network Development and MARC Standards Office. Functional analysis of the MARC 21 bibliographic and holdings formats. http://www.loc.gov/marc/marc-functional-analysis.

Adding User-Editing to a Catalogue of Cartoon Drawings

John Bovey

Computing Laboratory, The University, Canterbury, Kent CT2 7NF, UK
j.d.bovey@kent.ac.uk

Abstract. This paper describes an ongoing project to enable user-editing on an existing online database of about 120,000 British newspaper cartoons at the University of Kent. It describes the cartoon catalogue itself and then describes how the online search website has been extended to allow users to edit catalogue records in a way that should be both safe and economical. Finally, it discusses the next stage of the project, which is to experiment with ways to encourage users to become contributors.

1 Introduction

Online databases are usually one-way in that their data is provided and updated centrally, and is consumed by users. This is in contrast to shared resources such as, for example, the Wikipedia [2],[3],[4], where the information it contains is entered and maintained by users. It must be true, though, that for many one-way information resources, the users have a lot of information that could potentially be added to the resource and would improve it for everyone if that could be done. This user information could be simple corrections or more substantial additions. If we want to capture this information then we need to find a way to encourage users to contribute their knowledge while protecting the existing data from damage. This is what we are trying to do with our own online database of newspaper cartoons. . The system can be seen at opal.kent.ac.uk/cartoonx-cgi/ccc.py.

2 The Cartoon Catalogue

The University of Kent Centre for the Study of Cartoons and Caricature (more usually just called the Cartoon Centre) was established in 1973 when the university became the custodian of a collection of original artwork of 20th Century British newspaper cartoons. The collection, which included large bodies of work by, among others, David Low, Sidney George Strube and Victor Weisz (Vicky), was recognized as being a useful historical research resource and so the Cartoon Centre was set up and work started on cataloguing. The original catalogue was based on cards and photographs, but eventually it was replaced by an online catalogue with digitized images. From 1998, the Cartoon Catalogue has been freely accessible as a web site [1]. The presence of digitized images makes the catalogue into a self-contained research resource.

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 457 – 460, 2006.

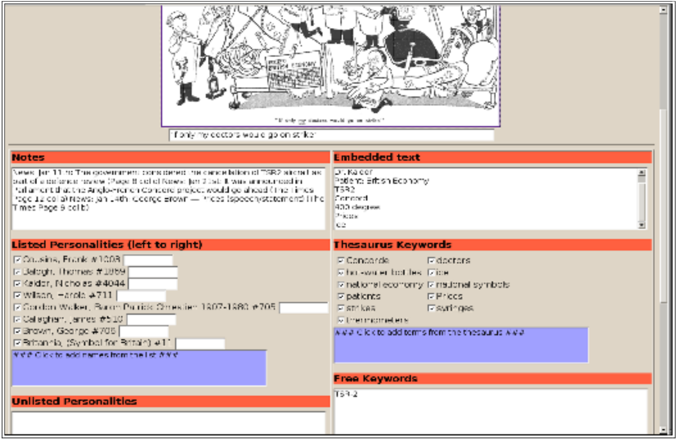
Each cartoon has a catalogue record and one or more digital images. The records are subdivided into fields, including mechanically derivable ones like the date and place of publication, the name of the artist, the caption and any text that can be transcribed from the drawing (from speech-bubbles, for example) as well as manually added subject terms and the names of depicted personalities.

There are strong reasons why the cartoon catalogue is a good candidate for user input One is that they are accessible public objects. Most of them appeared in newspapers or magazines and commented on events that were public knowledge at the time. No esoteric knowledge or special skills are needed to put a cartoon into context, just information that would have been widely available when the cartoon was published. Another feature of cartoons that makes them promising is that many are relevant to subjects that have their own interest groups. Cartoons on historical events (wars or elections, for example) could have information added by experts on the period when they occurred. Cartoons about sport could be put in contexts by knowledgeable sports enthusiasts, and so on. Cartoons are used as background by school teachers and could have their cataloguing improved as part of that process.

3 How User Contribution Works

A key feature of this project is that user edits are moderated by a member of the Cartoon Centre staff, but that this moderating should not be labor-intensive for staff or intrusive for contributors. The catalogue has a conventional search interface which allows the user to type in search statements and displays pages of summaries that can, in turn, be selected to see full catalogue records A user who wants to add information to a record or correct an error clicks on the Edit this record link in the top right hand corner of the record. The record is then redisplayed as a web form with the meta-data in editable boxes. Any of these fields can be edited or added-to, and the modified record saved back to the catalogue. Changes that have been made to a record but not yet moderated are visible to the user who made them so that they can be checked, corrected or extended. Other users, however, cannot see the changes until they have been accepted by the moderator.

The moderator's main web page contains a list showing a summary of outstanding edits that are waiting to be approved or rejected. Each line in the list shows which cartoon has been edited, the fields involved, the number of words changed, and also has information about the editing history of the contributor who made the edits. In particular, it shows the number of edits they have had rejected and accepted, and a credibility score for the contributor based on their past editing record. This list can be sorted by any of these fields by clicking on the head of the column. This means that a moderator who wants to deal with the small edits first can bring them to the top of the list. Similarly, the edits by high-credibility contributors can be brought to the top of the list and dealt with quickly.



"If only my doctors would go on strike"

Notes

News: Jan 11th: The government considered the cancellation of TSR2 aircraft as part of a defence review (Page 8 col e) News: Jan 21st: It was announced in Parliament that the Anglo-French Concorde project would go ahead (The Times Page 12 col a) News: Jan 14th: George Brown — Prices (speech/statement) (The Times Page 9 col b)

Embedded text

Dr. Kaldor
Patient: British Economy
TSR2
Concord
400 degrees
Prices
62

Listed Personalities (left to right)

☑ Cousins, Frank #1008
☑ Balogh, Thomas #1869
☑ Kaldor, Nicholas #4044
☑ Wilson, Harold #711
☑ Gordon Walker, Baron Patrick Chrestien 1907-1980 #705
☑ Callaghan, James #510
☑ Brown, George #706
☑ Britannia, (Symbol for Britain) #17

Click to add names from the list

Thesaurus Keywords

☑ Concorde ☑ doctors
☑ hot-water bottles ☑ ice
☑ national economy ☑ national symbols
☑ patients ☑ Prices
☑ strikes ☑ syringes
☑ thermometers

Click to add terms from the thesaurus

Free Keywords

TSR-2

Unlisted Personalities

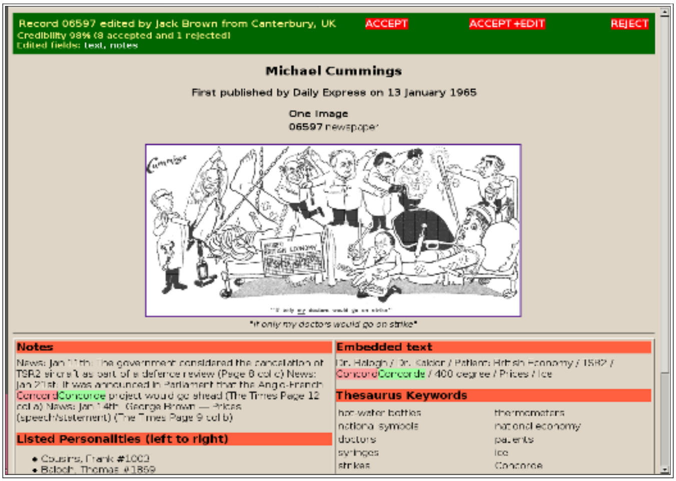
Fig. 1. The cartoon record displayed for editing

The page that the moderator uses to check individual cartoons is shown in Fig 2. The changed words are highlighted, with deletions in red and additions in green. At the top are three links: one to accept the edits as they are, one to accept them with minor changes by the moderator, and one to reject the edits outright. Clicking on the reject link puts up a web form that can be used to send an email to the contributor explaining why their edits were rejected.

4 Implementation

The original Cartoon Centre catalogue software was developed in-house because there were no appropriate products available in 1989. This was also the case when we first put the catalogue on the web and is true of the system described in this paper. The system runs on a Linux PC and uses the Apache web server. The retrieval and editing software is written in Python, with just the time-critical parts written in C. We also use a MySQL database to store the thesaurus, name-list, contributor details and to keep track of records that have been edited but not yet moderated. The retrieval software uses inverted files and a B-tree index to provide fast searching of Boolean queries.

The catalogue records are stored in individual XML files. Each record can have one or more edit elements and one or more history elements. Each edit element represents an editing session by a contributor and contains pending changes. It is also tagged with the identification number of the contributor and the date and time when the edits were submitted. The history elements are similar, but they represent edits that have been accepted and incorporated in the main record. This means that each record contains a history of all the changes that have been made to it, when and by whom. This information can be used to generate credits to contributors or, if something goes wrong, to unwind all the edits by an individual contributor.



Record 06597 edited by Jack Brown from Canterbury, UK
Credibility 98% (8 accepted and 1 rejected)
Edited fields: text, notes

Michael Cummings

First published by Daily Express on 13 January 1965

One Image
06597 newspaper

"If only my doctors would go on strike"

Notes

News: Jan 11th: The government considered the cancelation of TSR2 aircraft as part of a defence review (Page 8 col c) News: Jan 21st: It was announced in Parliament that the Anglo-French ConcordConcorde project would go ahead (The Times Page 12 col a) News: Jan 14th: George Brown — Prices (speech/statement) (The Times Page 9 col b)

Listed Personalities (left to right)

- Cousins, Frank #1003
- Balogh, Thomas #1859

Embedded text

Dr. Balogh / Dr. Kaldor / Patient: British Economy / TSR2 / ConcordConcorde / 400 degree / Prices / Ice

Thesaurus Keywords

hot-water bottles	thermometers
national symbols	national economy
doctors	patients
syringes	ice
strikes	Concorde

Fig. 2. The edited record as seen by the moderator

5 Current Status and Future Work

The Cartoon Catalogue has been running with user-editing for about two months and any user who takes a few minutes to register can become a contributor and edit the catalogue records, with any edits being quickly checked and incorporated into the main catalogue. Each record includes a full editing history that allows changes to be easily reversed or, more likely, acknowledged. The next stage of the project is to find ways to encourage and motivate potential editors. When we started the project, we did not know whether the major problem would be controlling over-enthusiastic contributors, protecting the catalogue against malicious edits, or encouraging people to contribute at all. It is now becoming clear that the main problem, at least initially, is encouraging people to contribute. This may be quite a challenge but we do at least now have a good test-bed for trying out different ways to encourage consumers of online data to become contributors.

References

1. Bovey, J.D., Providing Web access to a catalogue of British newspaper cartoons. Program. 37 (2003) 16-24
2. Leuf, B and Cunningham, W., The Wiki Way: Addison Wesley (2001)
3. Viegas, F.B., Wattenberg, M., Dave, K., Studying Cooperation and Conflict between Authors with History Flow Visualisations: Computer Human Interfaces Conference, Vienna, 2004 (available from web.media.mit.edu/~fviegas/papers/history_flow.pdf)
4. Wikipedia, the free encyclopedia: Available at en.wikipedia.org/wiki/Main_Page.

ALVIS - Superpeer Semantic Search Engine - ECDL 2006 Demo Submission

Gert Schmeltz Pedersen[1], Anders Ardö[2], Marc Cromme[3], Mike Taylor[4], and Wray Buntine[5]

[1] Technical University Library of Denmark
[2] Lund University
[3] Index Data ApS
[4] Index Data UK
[5] Helsinki University of Technology
gsp@dtv.dk, anders@it.lth.se, marc@indexdata.dk, mike@indexdata.com, buntine@hiit.fi

1 Introduction

ALVIS is a European project (IST-1-002068-STP) building a semantic-based peer-to-peer search engine. A consortium of eleven partners from six European Community countries, Finland, France, Sweden, Denmark, Spain, and Slovenia, plus Switzerland and China, contribute expertise in a broad range of specialities including network topologies, routing algorithms, probabilistic approaches to information retrieval, linguistic analysis and bioinformatics. The project runs from 1 January 2004 to 31 December 2006. Pointers to scientific papers and download sites for components can be found at http://www.alvis.info/.

2 Research Problem

The ALVIS project aims at developing an open source search engine. Since global search engines require vast scale and resources, the project has instead chosen to focus on a modest self-sustaining development path. To this end, two complementary activities are being pursued and integrated: open tools for building high-quality topic-specific search engines, and an infrastructure for supporting a peer-to-peer network. The topic-specific engines will provide some semantic awareness using information extraction technology, to make the system attractive for our initial targeted user base: professional associations, academic groups and digital libraries. The peer-to-peer infrastructure will allow inter-operability, the sharing of search resources, and the distributed solution of search tasks.

The objectives for the ALVIS project are as follows:

1. Subject area specialists will be able to run a small topic-specific search engine, with superior quality over general search engines.
2. The technology will have been developed so that when enough of these engines are installed, they can be tied together for general access through a distributed protocol, so that users can obtain search services without having to know the individual sites.

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 461–462, 2006.

3. Adequate scientific studies are made to provide some assurance that the system can operate and scale effectively and efficiently.

3 Methodology and Techniques

The major components of the finished system are:

- Document system: document acquisition by several methods, and document processing, doing linguistic and semantic enrichment, called the document processing pipeline.
- Maintenance system: collection-wide analysis to develop linguistic and semantic resources.
- Runtime system: indexing and query handling at a local peer.
- Peer-to-peer system: distributed query handling.

Focused web crawling is the major document acquisition method in ALVIS. It is capable of generating topic-specific databases of web pages by crawling the web and only saving relevant (i.e. topic-specific) pages. The system for automatic subject classification, which determines topical relevance and keeps the crawler focused, is based on matching of terms from a topic definition with the text of the document to be classified. The topic definition forms a hierarchical classification system of topic classes with controlled thesaurus terms associated with each class.

Relevant documents are further preprocessed and normalized by cleaning the HTML code, converting the character set to UTF-8 and identifying the language.

The document processing pipeline is based on XML metadata formats that allow a document to acquire successive layers of enrichment, and allow different kinds of programs to participate in the ALVIS system from different network locations. A document input to the system is converted to the ALVIS XML format and then enriched with acquisition data, derived structural, semantic and linguistic data for subsequent indexing, display and relevance calculations.

In the development of the user interface for the runtime system, we explore the possibilities that arise from the new information available about documents, and the new search modes made available due to the semantic and linguistic processing.

The peer-to-peer system appears as a single, simple proxy to both search user clients and document servers. Each peer dynamically maintains a pool of ”neighbour” peers that it knows about. A query is routed to the most relevant neighbours based on the topics areas that they cover. Semantic alignment of query interpretation between peers is facilitated by the use of ”context sets”, which rigorously define the semantics of the indexes they provide for use in queries.

4 The Demonstration

The intended demonstration is based on a workbench, where superpeers are formed by linking the major components by a document pipeline, and where a number of peers are cooperating on distributed query handling. The demonstration will highlight the principles of the major components and will use topic-specific test data sets, from bioinformatics and materials science.

Beyond Error Tolerance: Finding Thematic Similarities in Music Digital Libraries

Tamar Berman[1], J. Stephen Downie[2], and Bart Berman[3]

[1]Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
Champaign, IL 61820, USA
tamar@uiuc.edu
[2]Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
Champaign, IL 61820, USA
jdownie@uiuc.edu
[3]Independent Researcher, http://www.notesonfranzschubert.com
P.O. Box 476, Tel Mond 40650, Israel
bart@berman.tk

Abstract. Current Music Information Retrieval (MIR) systems focus on melody based retrieval with some tolerance for user errors in the melody specification. The system described here presents a novel method for theme retrieval: A theme is described as a list of musical events, containing melody and harmony features, which must be presented in a given order and within a given time frame. The system retrieves musical phrases that fit the description. A system of this type could serve musicians and listeners who wish to discover thematically similar phrases in music digital libraries. The prototype and underlying model have been tested on midi sequences of music by W.A. Mozart and have shown good performance results.

1 Introduction

Many musicians are familiar with the classic Dictionary of Musical Themes by Harold Barlow and Sam Morgenstern [1]. The dictionary includes about 10,000 themes present in classical instrumental pieces. It contains a unique notation index, in which a theme can be looked up based on a sequence of letters which represent the first few pitches of the theme's melody, transposed to the key of C. Creating an electronic implementation of this is a fairly straightforward task, and indeed such an implementation exists on the web [2]. Implementations which would be tolerant of user errors such as insertion of an incorrect pitch or omission of a required pitch in the theme description are well within the range of current Music Information Retrieval (MIR) technologies: n-grams [3], Markov models [4] and string matching techniques [5] have all been applied successfully to this task. However, in many cases themes are not defined by melody alone: often, the harmony is the primary describer of the theme. This project presents a novel access method: here, the user specifies the theme schematically, as a sequence of melody and harmony events that must be presented in a

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 463 – 466, 2006.

given order and within a given time frame. The system retrieves musical phrases that fit this description from a corpus of midi files. The retrieved phrases will often be reminiscent of each other, though not identical. This type of retrieval could serve musicians and listeners who wish to discover thematically similar phrases in music digital libraries.

2 Example

Figure 1 shows the first theme in the Allegro of Mozart's Clarinet Concerto in A, K622, as it appears in the Barlow and Morgenstern dictionary.

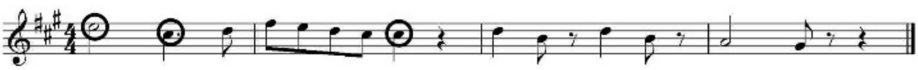


Fig. 1. First theme in Allegro of Mozart's Clarinet Concerto in A, K622

This theme could be described as the melodic sequence {E C# D F# E D C# C# D B D B A G#}, or transposed to C as GEFAGFEEFDFDCB. Indeed, the latter sequence is the one used in the B&M dictionary as the search key for this theme. An alternative description that uses a rhythmic encoding could be {Half, Dotted Quarter, Eighth, Eighth, Eighth, Eighth, Eighth, Quarter}. This prototype focuses on a third option, which encodes the theme as a sequence of harmony events. A musician described the core harmony of the theme as a sequence consisting of the following three events:

First event: A, C#, E with E as top voice
Second event: A, C# with C# as top voice
Third event: A, C#

In response to this specification, the system retrieved instances of the theme as they are presented in the Allegro. Examples of these are shown in Figures 2 and 3.

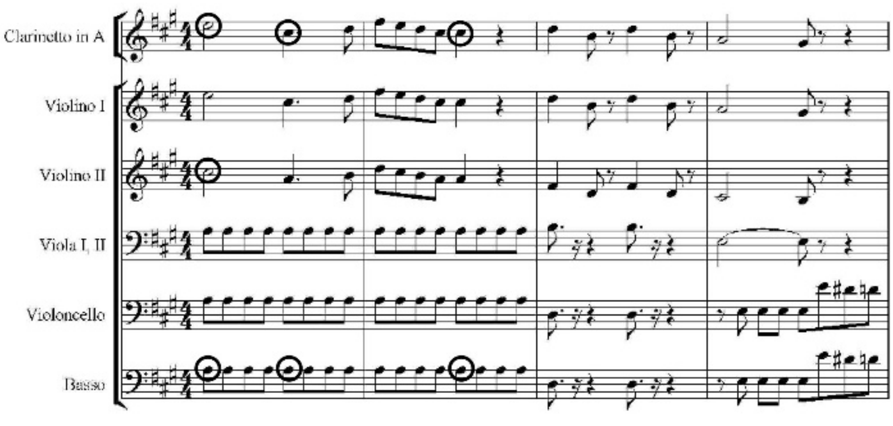


Fig. 2. Mozart Clarinet Concerto in A, K622, Allegro, measures 1-4

Figure 2 shows the first four measures in the Allegro of Mozart's Clarinet Concerto in A, K622. This is the first presentation of this theme. All of the theme's typical features are present here, and therefore any of the retrieval methods cited above – by melody, by rhythm or by harmony – would have yielded a successful retrieval of this instance. Figure 3, however, shows a somewhat transformed version of the theme. Neither the melody key {E C# D F# E D C# C# D B D B A G#} nor the rhythm key {Half, Dotted Quarter, Eighth, Eighth, Eighth, Eighth, Eighth, Quarter} would have succeeded in retrieving this instance, yet it was retrieved successfully by the harmony specification described above.

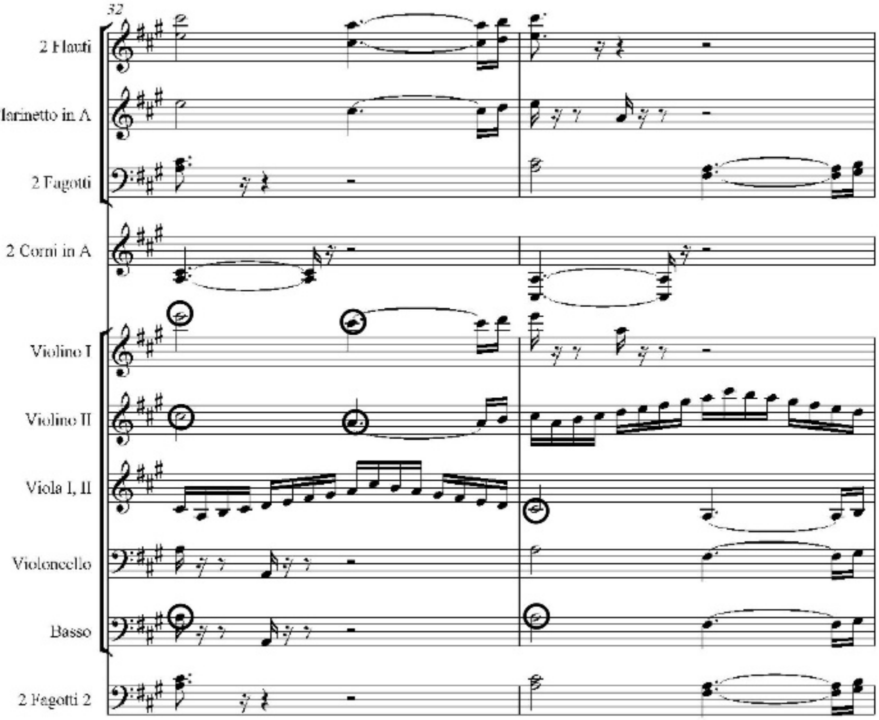


Fig. 3. Mozart Clarinet Concerto in A, K622, Allegro, measures 32-33[1]

3 The Model

In this model, music is conceived as an equally-spaced time series of 12-dimensional vectors. Each element in the time series, called a harmonic window, describes the pitch content of the time interval contained in the window. For example, a harmonic window which starts 5 seconds into the piece and ends 6 seconds into the piece describes, for each pitch class, its role within that time frame: top voice, bass, middle or

[1] Measure numbers refer to locations within midi files as they appear in classicalarchives.com

absent. The series are constructed on the basis of two parameters: window length and onset interval. The onset interval defines the time that elapses between window onsets and is somewhat analogous to the sampling rate used in audio files. The window length describes the amount of tolerance permitted for pitches to be considered as sounding simultaneously.

The model is distinctly different from traditional MIR systems in its definition of a sequence: In Query-By-Humming systems, if you are searching for "a b c", then "a b xx c" will score higher than "a yy b xx c", because xx and yy are viewed as errors. In this model, "a b c", "a b xx c" and "a yy b xx c" are equivalent if they complete within a given time frame, such as half a second. Additionally, each of the required items (a, b, c) can be a combination of pitches which are played close to each other.

4 Model Testing and Performance

The model has been tested on midi sequences of music by W.A. Mozart obtained from the Classical Music Archives [6]. Preliminary evaluation of the model's performance yields encouraging results: Straightforward, query based retrieval achieved up to 88% success rate for precision and up to 86% for recall. Ranking based retrieval using multiple criteria showed yet higher success, where nearly all of the instances produced by the system were rated as musically correct examples of the theme by the project's musicians.

Acknowledgements

This work is supported by National Science Foundation (NSF) grant IIS-0328471.

References

1. Barlow H., Morgenstern S. *A Dictionary of Musical Themes*, 1948
2. The Multimedia Library, http://www.multimedialibrary.com/barlow/
3. Downie J. S., Nelson M. "Evaluation of a Simple and Effective Music Information Retrieval Method" in 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece, 2000
4. Birmingham W. P., Dannenberg R. B., Wakefield G. H., Bartsch M., Bykowski D., Mazzoni D., Meek C., Mellody M., Rand W. "Musart: Music Retrieval via Aural Queries", ISMIR 2001, Bloomington, Indiana, 2001
5. McNab, R. J., Smith L.A., Witten I. H., Henderson C. L., Cunningham S. J. "Towards the Digital Music Library: Tune Retrieval from Acoustic Input" in *Digital Libraries*, 1996
6. Classical Music Archives, http://www.classicalarchives.com

Comparing and Combining Two Approaches to Automated Subject Classification of Text

Koraljka Golub[1], Anders Ardö[1], Dunja Mladenić[2], and Marko Grobelnik[2]

[1] KnowLib Research Group, Dept. of Information Technology, Lund University, Sweden
{Koraljka.Golub, Anders.Ardo}@it.lth.se
[2] J. Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
{Dunja.Mladenic, Marko.Grobelnik}@ijs.si

Abstract. A machine-learning and a string-matching approach to automated subject classification of text were compared, as to their performance, advantages and downsides. The former approach was based on an SVM algorithm, while the latter comprised string-matching between a controlled vocabulary and words in the text to be classified. Data collection consisted of a subset from Compendex, classified into six different classes. It was shown that SVM on average outperforms the string-matching approach: our hypothesis that SVM yields better recall and string-matching better precision was confirmed only on one of the classes. The two approaches being complementary, we investigated different combinations of the two based on combining their vocabularies. The results have shown that the original approaches, i.e. machine-learning approach without using background knowledge from the controlled vocabulary, and string-matching approach based on controlled vocabulary, outperform approaches in which combinations of automatically and manually obtained terms were used. Reasons for these results need further investigation, including a larger data collection and combining the two using predictions.

1 Methodology

The string-matching algorithm [1] classifies text documents into classes of the Ei classification scheme and thesaurus [2], based on simple string-matching between terms in the term list, derived from Ei, and terms in the document being classified. Ei contains several different types of terms and relationships, out of which we used captions, preferred and non-preferred terms. The term list (a model for classification) is organized as a list of triplets: term, class to which it maps, and weight. Each class in the original list is designated a number of term entries. No cut-off was used.

The second algorithm we used was linear SVM (support vector machine), a state-of-the art machine-learning algorithm, commonly used for text classification. We used binary SVM, the implementation from TextGarden [3]. We preprocessed the text by removing stop-words and representing each document using the standard bag-of-words approach containing individual words, enriched by frequent phrases (occurring at least four times in the data collection). The frequent phrases containing up to five consecutive words were automatically generated, as described in [4]. The model was trained on a part of data collection leaving the other part to be

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 467 – 470, 2006.

classified using a standard approach of ten-fold cross validation. The binary classification model was automatically constructed for each of the six classes (see 2), taking all the training examples of the class as positive and all the other training examples as negative. Each example from the data collection was classified by each of the six models. For each example, we report all the classes that are above the threshold of zero.

2 Experimental Setting

Data collection consisted of a subset of paper records from the Compendex database [5], classified into six selected classes. Each document can belong to more than one class. Fields of the records that were used to classify are title, abstract and uncontrolled terms in the string-matching algorithm, and title and abstract in SVM.

In this first run of the experiment, only the six classes were selected in order to provide us with indications for further possibilities. Classes 723.1.1 (Computer Programming Languages), 723.4 (Artificial Intelligence), and 903.3 (Information Retrieval and Use) each had 4400 examples (the maximum allowed by the Compendex database provider), 722.3 (Data Communication Equipment and Techniques) 2800, 402 (Buildings and Towers) 4283, and 903 (Information Science) 3823 examples.

The linear SVM in the original setting was trained with no feature selection except the stop-word removal. Additionally, three experiments were conducted using feature selection, taking:

1. only the terms that are present in the controlled vocabulary;
2. the top 1000 terms from centroid tf-idf vectors for each class (terms that are characteristic for the class – descriptive terms);
3. the top 1000 terms from the SVM-normal trained on a binary classification problem for each class (terms that distinguish one class form the rest – distinctive terms).

In the experiments with string-matching algorithm, four different term lists were created, and we report performance for each of them:

1. the original one, based on the controlled vocabulary;
2. the one based on automatically extracted descriptive keywords from the documents belonging to their classes;
3. the one based on automatically extracted distinctive keywords from the documents belonging to their classes;
4. the one based on union of the first and the second list.

In lists 2, 3, and 4, the same number of keywords was assigned per class as in the original one.

Evaluation was based on comparing automatically assigned classes against the intellectually assigned classes given in the data collection. Precision (Prec.), recall (Rec.) and F1 measure were used as standard evaluation measures. Both standard ways of calculating the average performance were used: macroaverage (macro.) and microaverage (micro.)

3 Experimental Results

We have experimentally compared performance of the two algorithms on our data in order to test two hypotheses both based on the observation that the two algorithms are complementary. Our first hypothesis was that, as the string-matching algorithm uses manually constructed model, we expect it to have higher precision than the machine learning with its automatically constructed model. On the other hand, while the machine-learning algorithm builds the model from the training data, we expect it to have higher recall in addition to being more flexible to changes in the data. Experiments have confirmed the hypothesis only on one of the six classes. Experimental results of the string-matching approach and the machine learning (SVM) approach (both using their original setting) are given in Table 1: SVM on average outperforms the string-matching algorithm. Different results were gained for different classes. The best results are for class 402, which we attribute to the fact that it has the highest number of term entries designating it (423). Class 903.3, on the other hand, has only 26 different term entries designating it in the string-matching term list, but string-matching largely outperforms SVM in precision (0.97 vs. 0.79). This is subject to further investigation.

Table 1. Experimental results comparing performance of the two approaches, and number of original terms per class (Terms). We can see that SVM performs better in all but one classes.

Class	**String-matching (SM)**				**Machine learning (SVM)**		
	Terms	**Rec.**	**Prec.**	**F1**	**Rec.**	**Prec.**	**F1**
402	423	0.58	0.49	0.53	**0.93**	**0.91**	0.92
722.3	292	0.12	0.26	0.16	**0.76**	**0.79**	0.78
723.1.1	137	0.34	0.32	0.33	**0.74**	**0.79**	0.76
723.4	61	0.37	0.39	0.38	**0.65**	**0.81**	0.72
903	58	0.28	0.61	0.38	**0.72**	**0.74**	0.73
903.3	26	0.32	**0.97**	0.48	**0.74**	0.79	0.76
Micro.		0.35	0.45	0.39	**0.78**	**0.81**	**0.78**
Macro.		0.34	0.51	0.38	**0.76**	**0.81**	**0.78**

The second hypothesis was that combining the two approaches via combining their vocabularies will result in improved performance. This hypothesis was not confirmed: both approaches have the best performance in the original setting (see Table 2). We attribute that to a large overlap between the controlled vocabulary and the document vocabulary that enables SVM to find the right terms for a good quality model.

Table 2. Macroavergaed experimental results comparing performance of SVM and string-matching approach (SM). We can see that both perform best using the original vocabularies.

				SVM – controlled		
	Rec.	**Prec.**	**F1**	**Rec.**	**Prec.**	**F1**
SVM – original (complete)	**0.76**	**0.81**	**0.78**	0.55	0.57	0.55
	SVM – descriptive			**SVM – distinctive**		
	Rec.	**Prec.**	**F1**	**Rec.**	**Prec.**	**F1**
Macroavg.	0.72	0.79	0.75	0.75	0.64	0.69
	SM - original (controlled)			**SM – distinctive + controlled**		
	Rec.	**Prec.**	**F1**	**Rec.**	**Prec.**	**F1**
Macroavg.	**0.55**	**0.68**	**0.61**	0.99	0.19	0.32
	SM– descriptive			**SM – distinctive**		
	Rec.	**Prec.**	**F1**	**Rec.**	**Prec.**	**F1**
Macroavg.	0.92	0.29	0.43	0.99	0.19	0.32

From Table 3 we can see that the string-matching algorithm the performance decreases due to a large drop in precision. Actually, almost every document gets all the six classes assigned, which increases recall to almost 100%. There is a possibility that low precision could be improved by introducing a cut-off value.

Acknowledgements

This work was supported by the Slovenian Research Agency and the IST Programme of the European Community under ALVIS (IST-1-002068-STP).

References

1. Golub, K. 2006. Automated subject classification of textual Web pages, based on a controlled vocabulary: challenges and recommendations. New review of hypermedia and multimedia, Special issue on knowledge organization systems and services, 2006(1).
2. Ei thesaurus, edited by J. Milstead, Engineering Information, Castle Point on the Hudson Hoboken, 1995. 2nd ed.
3. Grobelnik, M., Mladenic, D. Text Mining Recipes, Springer-Verlag, Berlin; Heidelberg; New York, 2006, accompanying software available at http://www.textmining.net.
4. Mladenic, D., Grobelnik, M. Feature selection on hierarchy of web documents. Journal of Decision Support Systems, 35, 45-87, 2003.
5. Compendex database. http://www.engineeringvillage2.org/.

Concept Space Interchange Protocol: A Protocol for Concept Map Based Resource Discovery in Educational Digital Libraries

Faisal Ahmad, Qianyi Gu, and Tamara Sumner

BOulder Learning Technologies Lab (BOLT), Dept. of Computer Science,
University of Colorado at Boulder, CO-80309, USA
{Faisal.Ahmad, Qianyi.Gu, Tamara.Sumner}@colorado.edu

Abstract. The Strand Map Service provides resource discovery in digital libraries using strand maps developed by the American Association for the Advancement of Science, project 2061. Strand maps are a special kind of concept maps that contains interconnected learning goals organized along grade groups and topical strands. The Strand Map Service provides programmatic access to AAAS strand maps that can be used by educational digital libraries to dynamically build resource discovery interfaces. The programmatic access to strand maps is enabled by the Concept Space Interchange Protocol, which provide following services (1) service capability determination, (2) resource alignment, and (3) search and retrieval of dynamically generated strand maps. The protocol is implemented as a web service and integration experiments have been performed for two educational digital libraries. In this poster we describe the Concept Space Interchange Protocol and its integration with educational digital libraries.

1 Introduction

In the world of ever increasing information access there is a growing need to disseminate, organize and process information as efficiently as possible. Effective information dissemination has been made possible by technological advances such as the internet, however, the problem of effective information organization and information processing is still unresolved. Concept maps present a possible answer to the unresolved issues of information organization and processing. Concept maps put special constraints on the information representation and visualization, forming the basis of a coherent and interrelated view of information. Traditionally concept maps are represented as node-arc diagrams. The nodes represent concepts and the arcs relate nodes by a set of relationships. This representation forms a web of information in which the relationship of each concept is well understood in relation to another concept.

Concept maps based instructional improvement can be seen as part of bigger science educational reform that has been the focus of American education reform over the last decade. As a part of this reform movement, the American Association for the Advancement of Science (AAAS) has formulated a set of learning goals called benchmarks organized into strand maps. Each strand map has a focus topic and contains related benchmarks. The benchmarks in a strand map are organized along grade

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 471 – 476, 2006.

groups and finer topics called strands. This two dimensional organization of benchmarks shows an increasing learner knowledge with age. AAAS has articulated 854 benchmark for science literacy for K-12 grades that are organized in approximately one hundred strand maps published in two volumes of Atlas of Science Literacy [1]. The strand maps provide instruction and learning paths for different science topics and can be used by educators and learners in a variety of manners such as facilitating learning, instructions planning, and as a framework for developing coherent curriculum.

The Strand Map Service (SMS) research project [2] uses the AAAS strand maps, and makes them available to digital libraries (DL) for resource discovery and navigation. By using this service, DL patrons can see the interrelated nature of concepts, represented as benchmarks and their relationships, and trivially find the resources suitable for teaching those concepts.

2 Concept Space Interchange Protocol

The Concept Space Interchange Protocol (CSIP) is the primary mode of interaction between digital libraries and the Strand Map Service. Its design is based on the REpresentational State Transfer (REST) web architecture style. REST is a document centric web service architecture style where each service request has a unique URL and each response is considered to be a transfer of representation of a document. CSIP provides four services that can be used by DLs to access strand maps information. The following services are provided by CSIP: service description, register query, submit resource, and query. (1) Service description request is used to dynamically determine the capabilities of the CSIP server. It can be initiated by using following URL through an HTTP Get method:

```
servername/ServiceDescription
```

The information is returned in XML format and includes version information, supported return formats, supported operators and other descriptive information about SMS server capabilities. The client DLs can tailor the services offered to patrons based on the response of this request. (2) Register query service request can be used by DLs to register their query format with SMS. This enables DLs to retrieve strand map information along with pre-constructed query string that can be used for resource discovery. This service can be initiated by using an HTTP Get or Post method and the following URL:

```
servername/RegisterQuery
```

(3) Submit resource service request is used as a means of community participation for the continual improvement of SMS. Resources that the client perceives to be ideal for teaching a benchmark can be submitted using this request. This resource goes through a quality assurance process and is added to the pool of exemplary resources maintained by SMS. This service can be initiated by using an HTTP Get method and the following URL:

```
servername/SubmitRequest?ObjectID=ID&Resource=URL&email=senderEmail
```

The ObjectID parameter in this request is the unique identification of a benchmark in the Strand Map Service metadata repository and the Resource parameter represents the URL of the exemplary resource. The email parameter is used for communication with the contributor of resource. (4) Query service request is used to get AAAS concept map information that can be used in DLs for resource discovery and navigation. This request can be made by using the HTTP Get or Post method with the following URL:

```
servername/Query
```

CSIP is partitioned into two parts: CSIP-core and CSIP-extension. This partitioning is done to support a 'spectrum of interoperability' to maximize its utility for a broad range of digital library projects [3]. The only difference between the CSIP-core and the CSIP-extension is the amount of query support available. CSIP-core supports a limited set of query constructs i.e. content type query. On the other hand, CSIP-extension also supports navigational query type. A content query search is similar to a text based query where the terms are matched against text present in metadata. The navigational query makes use of the rich relationships that are part of the AAAS strand maps such as *contributes to achieving*, *is part of* etc. The navigational query starts from one benchmark and find all object that are related to it by a specific relation. An example of navigational query is to find all the neighbors or prerequisites of a benchmark.

CSIP Query language has a number of features that enhance its facility and expressiveness. Facility of a language is defined as "*the degree to which programs in a language are easy to write, not in the sense of physical effort but in the sense of mental effort*"[4]. Facility of a language is important for increasing its usability. Query construction is a complex task because it has to satisfy a given requirement, dictated by the user interface, under the constraints of the strand map representation. To ease query construction and to improve CSIP facility, the query language has constructs that parallel the strand maps ideas, components and constructs. Therefore, it becomes easy to translate strand maps related concepts, knowledge and search requirements to query constructs for query construction purposes. The other important aspect of CSIP query language is its expressiveness. Expressiveness is defined as "*A language is said to satisfy expressiveness criterion for a problem if there is a program in the language that solves the problem*" [4]. Expressiveness of a language determines its utility and usefulness. The CSIP query language constructs can be combined in number of ways to get the same information. This flexibility of query construction allows for dealing with different scenarios and context of use.

3 Sample SMS Interface

Given the features of CSIP-query language, it can be used effectively by educational digital libraries for supporting resource discovery and navigation based on strand maps. We have built an SMS interface for Nederland High school library web-site shown in figure 1 & 2. The entry page, figure 1, shows the list of available strand maps in iconic form along with the topics covered in each of them.

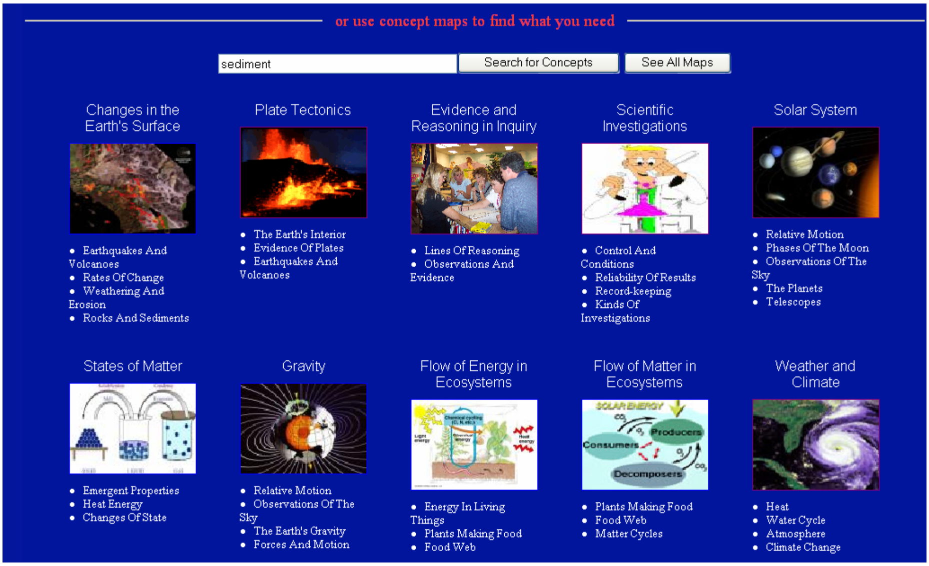
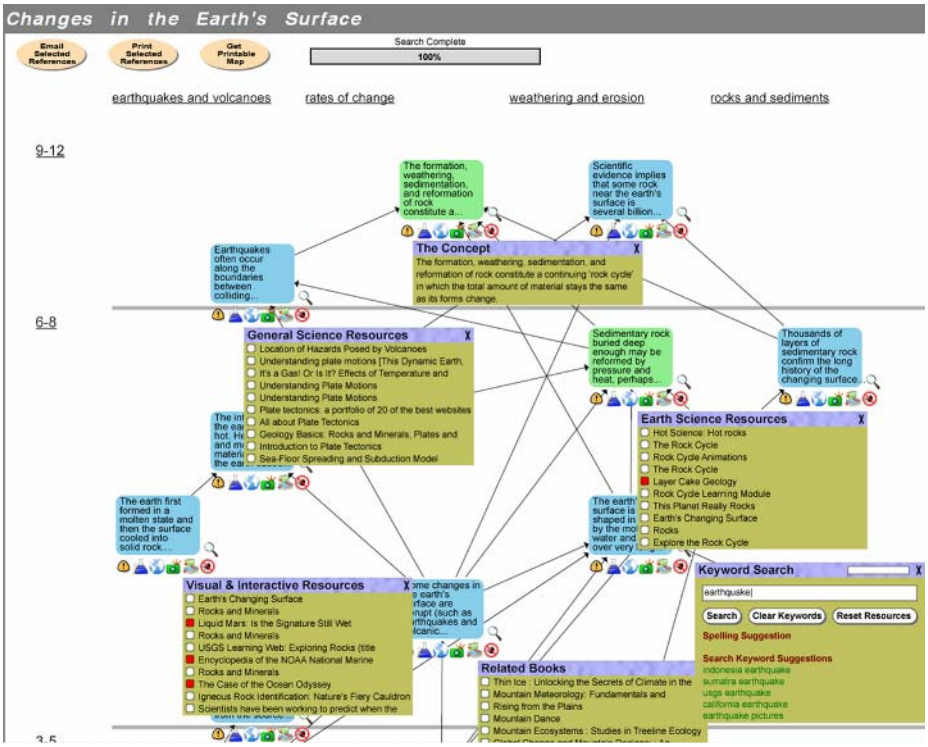


Fig. 1. User can search for concepts using the search box. Clicking on one of the icons bring up the appropriate strand map along with resources useful for teaching benchmarks.

For example, the *Evidence and Reasoning in Inquiry* strand map covers the *Lines of Reasoning* and *Observations and Evidence* sub-topics. A teacher can click on one of these strand map icons to explore the concepts in the map and associated resources. But sometimes teachers find it difficult to locate a particular topic of interest just by browsing the map information. In this scenario teachers can type the desired keyword in the search box, shown at the top of figure 1, and find the matching concepts. The results (not shown in the figure) display strand map icons in the order of decreasing number of matching benchmarks. For example, if a teacher performed a search for the keyword *sediment*, she will only see the *Changes in Earth's Surface* strand map because this is the only map with benchmarks that contain the desired keyword. Once the teacher clicks on *Changes in Earth's Surface* map she sees the annotated strand map as shown in figure 2. The title of the strand map is displayed at the top left corner. The blue boxes contain brief description of a benchmark and the arrows interconnect these boxes with meaningful relationships. The green boxes show brief descriptions of the benchmarks with matching keywords (i.e. *sediment*). The strand names are shown at the top of the map and the grade ranges are shown at the left edge of the map. The horizontal grey lines define the grade boundaries for the benchmarks. The small icons at the bottom of each benchmark box provide additional information and supporting educational resources available in digital libraries as well as commercial web-site (i.e. National Science Digital Library (NSDL), Digital Library for Earth System Education (DLESE), Amazon and Yahoo). Clicking on the *exclamation icon* (⚠) brings the detailed description of the benchmark, labeled as *The Concept* in the title of the popup window. Once the teacher is done reading the benchmark details she can click on the *beaker icon* (🧪) to see general science resources available for the given benchmark. She can click on the resource titles to open up the resource in a new browser window. If the teacher likes the resource she can bookmark it by clicking on the box next to the resource title in the popup window. In addition, clicking on the title of the popup window takes the teacher to the NSDL which allows for deeper inspection of general science resources. Similarly, the teacher can click on: *earth icon* (🌍) to see earth science resources, *camera icon* (📷) to see visual & interactive resources, book icon (📚) to see books relevant to teaching the benchmark, and the *do-not-enter icon* (⛔) to see search results from the World Wide Web. If the teacher is not satisfied with the quality or suitability of the resources she can click on the *search-in-context icon* (🔍) to open a search popup window. Typing a new keyword and clicking the search button, within the search popup window, changes the resources that appear in each of the icons for this benchmark. The search popup window also provides support for spelling corrections and keyword suggestions to facilitate the search process. Finally the buttons at the top left corner of Figure 1 allow teachers to email the bookmarked resources, print the bookmarked resources and print the strand map for annotation and off-line use in classroom or for discussions with peers. This strand map interface shields NHS teachers from the complexities associated with using multiple digital libraries and searching out of context.


Changes in the Earth's Surface
Email Selected References
Print Selected References
Get Printable Map
Search Complete
100%
earthquakes and volcanoes
rates of change
weathering and erosion
rocks and sediments
9-12
6-8
3-5
The formation, weathering, sedimentation, and reformation of rock constitute a...
Scientific evidence implies that some rock near the earth's surface is several billion...
Earthquakes often occur along the boundaries between colliding...
Sedimentary rock buried deep enough may be reformed by pressure and heat, perhaps...
Thousands of layers of sedimentary rock confirm the long history of the changing surface...
The in the ea hot. H and m materi the ea
The earth first formed in a molten state and then the surface cooled into solid rock....
The earth' surface is shaped in by the mo water and over very
ome changes in e earth's rface are rupt (such as rthquakes and lcanic...

The Concept X

The formation, weathering, sedimentation, and reformation of rock constitute a continuing 'rock cycle' in which the total amount of material stays the same as its forms change.

General Science Resources X

- ☐ Location of Hazards Posed by Volcanoes
- ☐ Understanding plate motions [This Dynamic Earth,
- ☐ It's a Gas! Or Is It? Effects of Temperature and
- ☐ Understanding Plate Motions
- ☐ Understanding Plate Motions
- ☐ Plate tectonics: a portfolio of 20 of the best websites
- ☐ All about Plate Tectonics
- ☐ Geology Basics: Rocks and Minerals, Plates and
- ☐ Introduction to Plate Tectonics
- ☐ Sea-Floor Spreading and Subduction Model

Earth Science Resources X

- ☐ Hot Science: Hot rocks
- ☐ The Rock Cycle
- ☐ Rock Cycle Animations
- ☐ The Rock Cycle
- ☒ Layer Cake Geology
- ☐ Rock Cycle Learning Module
- ☐ This Planet Really Rocks
- ☐ Earth's Changing Surface
- ☐ Rocks
- ☐ Explore the Rock Cycle

Visual & Interactive Resources X

- ☐ Earth's Changing Surface
- ☐ Rocks and Minerals
- ☒ Liquid Mars: Is the Signature Still Wet
- ☐ Rocks and Minerals
- ☐ USGS Learning Web: Exploring Rocks (title
- ☒ Encyclopedia of the NOAA National Marine
- ☐ Rocks and Minerals
- ☒ The Case of the Ocean Odyssey
- ☐ Igneous Rock Identification: Nature's Fiery Cauldron
- ☐ Scientists have been working to predict when the

Related Books

- ☐ Thin Ice : Unlocking the Secrets of Climate in the
- ☐ Mountain Meteorology: Fundamentals and
- ☐ Rising from the Plains
- ☐ Mountain Dance
- ☐ Mountain Ecosystems : Studies in Treeline Ecology

Keyword Search X

earthquake|

Search | Clear Keywords | Reset Resources

Spelling Suggestion

Search Keyword Suggestions

indonesia earthquake
sumatra earthquake
usgs earthquake
california earthquake
earthquake pictures

Fig. 2. The icons below the benchmarks show supporting resources from different digital libraries and commertial web-sites. The use can also initiaite search in context by clicking on magnifying glass icon.

4 Conclusion and Future Work

Concept Space Interchange Protocol expressiveness and alignment with strand map constructs makes it useful for concept maps information exchange. We believe Concept Space Interchange Protocol can be used with concept spaces other than AAAS strand maps with little or no modifications. Hence, CSIP provides a starting point of a uniform way of discovering digital library resources using concept map.

References

1. American Association for Advancement of Science, Project 2061, *Atlas of Science Literacy*, vol. 1: Washington, DC, 2001.
2. Sumner, T., Ahmad, F., Bhushan, S., Gu, Q., Molina, F., Willard, S., et al. (2005). *Linking learning goals and educational resources through interactive concept map visualizations*, International Journal on Digital Libraries, 5, 18-24.
3. Arms, W. Y., Lagoze, C., Krafft, D., Marisa, R., Saylor, J., Terrizzi, C., Van de Sompel, H., *A Spectrum of Interoperability, The Site for Science Prototype for the NSDL*, D-Lib Magazine, vol. 8, 2002.
4. Lewis, C., Rieman, J., Bell, B., *Problem-Centered Design for Expressiveness and Facility in a Graphical Programming System*, Technical report: CU-CS-479-90, University of Colorado at Boulder, 1990.

Design of a Cross-Media Indexing System

Murat Yakıcı and Fabio Crestani

i-lab group
Department of Computer and Information Sciences
University of Strathclyde
26 Richmond Street, Livingstone Tower, G1 1XH, Glasgow, UK
{murat.yakici, fabio.crestani}@cis.strath.ac.uk

Abstract. There is a lack of an integrated technology that will increase effective usage of the vast and heterogeneous multi-lingual and multimedia digital content. The need is being expressed insistently by end-users, and professionals in content business. The EU-IST Framework 6 Reveal-This (R-T) project aims at developing a *complete* and *integrated* content programming technology able to capture, semantically index, categorise, multimedia and multilingual digital content, whilst providing search, summarisation and translation functionalities. In order to fulfill this, the project proposes an architectural unit called Cross-Media Indexing Component (CMIC). CMIC leverages the individual potential of each indexing information generated by the analyzers of diverse modalities such as speech, text and image. It hypothesises that a system which combines and cross analyses different high-level modal descriptions of the same audio-visual content will perform better at retrieval and filtering time. The initial prototype utilises the *Multiple Evidence* approach by establishing links among the modality specific descriptions in order to depict topical similarity in the semantic textual space. This paper gives an overview of the project, CMIC's enrichment approach and its support for retrieval.

1 The Reveal-This project

The main objective of the R-T project is to design, develop and test a complete and integrated infrastructure that will allow the user to store, categorize and retrieve multimedia and multi-lingual digital content across different sources (such as TV, radio, Web). The project integrates a whole range of information access technologies across media and languages.

A major challenge lies in developing suitable cross-media[1] representations for the processes of *retrieval*, *categorization* and *summarization*. The project heavily emphasises on promoting cross modal analysis and indexing techniques to improve the effectiveness of the system. In order to overcome the challenge, R-T system relies on CMIC, which can combine and cross analyse modality specific high-level descriptions of an audio-visual content in a semantic textual space.

[1] Here, the term *media* refers to the format in which information on a topic is conveyed by one source (e.g. the audio, the image and the text of some video segment).

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 477–480, 2006.

2 Cross-Media Indexing

The process of cross-media indexing consists of building relationships among high-level features and concepts extracted from different modalities such as speech, image and text analysis. Given the use of advanced audio-visual content analysis technologies, a unique and complete description of the topical content can be reconciled.

CMIC is a standalone server which constitutes the middle layer of the Cross-Media Analysis and Indexing Subsystem (CAIS) (see Fig. 1)).

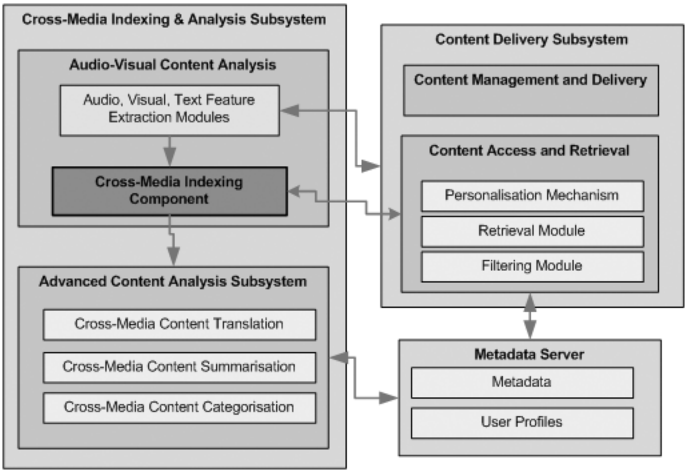

Cross-Media Indexing & Analysis Subsystem
Audio-Visual Content Analysis
Audio, Visual, Text Feature Extraction Modules
Cross-Media Indexing Component
Advanced Content Analysis Subsystem
Cross-Media Content Translation
Cross-Media Content Summarisation
Cross-Media Content Categorisation
Content Delivery Subsystem
Content Management and Delivery
Content Access and Retrieval
Personalisation Mechanism
Retrieval Module
Filtering Module
Metadata Server
Metadata
User Profiles

Fig. 1. The Reveal-This system architecture

CAIS facilitates the generation of a rich index which derives from identification of semantic concepts, entities and facts in addition to face and category information extracted by state-of-art speech, text, image processing units and related categorizers.

2.1 Process Model

First, each processing unit completes its analysis on a multimedia stream. During this step, segments are identified and regarded as indexable units. The features are gathered, aligned, synchronized and merged into an intermediate representation. CMIC is given this high-level feature and concept set for further processing. This process can be divided into *Analysis* and *Indexing phases*. The analysis phase comprises of parsing and transformation of an input stream, noise filtering and lexical analysis tasks. All modality specific descriptions are transformed and mapped to their corresponding MPEG-7 [1] elements. MPEG-7's extension mechanism is used to define new descriptors where necessary in order to meet the R-T system requirements. CMIC progressively enriches the given input by establishing cross links between each modal descriptions of the same segment. This enrichment is done by using an indexing model. As a result, CMIC produces a unified view of the content in MPEG-7 with a measure of uncertainty attached. Subsequently, this output is handed over to the other subsystems such as Cross-Lingual Translation, Cross-Media Summarisation and Content Delivery.

2.2 Indexing Model

The prototype utilises Dempster-Shafer's Theory of Evidence [2] approach for establishing links among the modality specific descriptions in order to depict *topical similarity* in the textual space. The theory has been extensively studied in image retrieval [3,4,5], structured document retrieval [6], but has never been applied in such a context.

Briefly, Dempster-Shafer combines two or more bodies of evidence defined with in the same *frame of discernment* T into one body of evidence. In our approach, a document d containing a term t's existence in one modality m is counted as an *evidence* to support the topical similarity hypothesis. Therefore, each modality is treated as a probability density function also called as *Base Probability Assignment (BPA).*

$$m(\emptyset)_d = 0 \quad \text{and} \quad 1 = \sum_{t \in d} m(\{t\}) \ . \tag{1}$$

where

$$\mathrm{m}_d(\{t\}) = \begin{cases} tf(d,t). \log_N(\frac{N}{n(t)}) & \text{if} \quad t \in d, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

$$m_d(T) = 1 - \sum_{t \in d} m_d(\{t\}) \ . \tag{3}$$

In (2), $tf(d,t)$ is calculated as the number of occurences of term t by the number of all terms in the document d. The logarithm $\log_N(\frac{N}{n(t)})$ is a variant of inverse document frequency where N represents the number of documents in the collection and $n(t)$ represents the number of documents that contain t. The preceding formula gracefully adheres to the theory by $\sum_{t \in d} m_d(\{t\}) \leq 1$ and therefore $m(T)$ is equal to the unassigned BPA (3) when $m_d(\{t\})$ is smaller than 1. We combine evidences from modality specific descriptions by applying the following combination rules from [6]:

$$m(\{t\}) = m_1 \otimes m_2(\{t\}) = \frac{1}{K}(m_1(\{t\}).m_2(\{t\}) + m_1(\{t\}).m_2(T) + m_2(\{t\}).m_1(T)) \ . \tag{4}$$

$$m(T) = m_1 \otimes m_2(T) = \frac{1}{K}(m_1(T).m_2(T)) \ . \tag{5}$$

where

$$\mathrm{K} = \left(\sum_t m_1(\{t\}).m_2(\{t\}) \right) + m_1(T).m_2(T) \ . \tag{6}$$

such that $m_1(\{t\}) > 0$ *and* $m_2(\{t\}) > 0$ conditions are satisfied.

3 Evaluation

In order to validate our arguments and thus the first prototype, we intend to evaluate the performance of different cross-media indexing models and the reliability of the single modality indexing modules. This will enable us to explore the pros and cons of various indexing models for the combination of evidence. In this context, we are currently pursuing two different strategies *known item search* and *task oriented user test*. Both strategies involve the construction of a test collection using real users which is still in progress. The collection that we build is a three-hour multimedia collection which covers politics, travel and news domains in English, French and Greek languages. It should also be noted that the final evaluation of the R-T system will be carried out by a user and task oriented approach involving home user and TV broadcast professionals.

4 Conclusions and Future Work

In this paper, we have given an overview of the work in progress which is part of the EU-IST Reveal-This project. Our approach to cross-media indexing was presented. We are on the way to show that the *Multiple Evidence* approach can be employed to robustly reconcile a unique and complete description of the topical content. The hypothesis of cross-media indexing and the models in use are still an open research area. At the moment, we are exploring other combination rules and indexing models.

References

1. Martínez, J.M.: MPEG-7: Overview of MPEG-7 description tools. IEEE Multimedia **9**(3) (2002) 83–93
2. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press (1976)
3. Jose, J.M., Harper, D.J.: A retrieval mechanism for semi-structured photographic collections. In: Proceedings of the DEXA, Springer-Verlag (1997)
4. Aslandogan, Y.A., Yu, C.T.: Multiple evidence combination in image retrieval: Diogenes searches for people on the web. In: Proceedings of the ACM SIGIR, Athens, Greece, ACM Press (2000)
5. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings of the ACM SIGIR, Toronto, Canada, ACM Press (2003)
6. Lalmas, M., Moutogianni, E.: A Dempster-Shafer indexing for the focussed retrieval of a hierarchically structured document space: Implementation and experiments on a web museum collection. In: Proceedings of RIAO, Paris, France (2000)

Desired Features of a News Aggregator Service: An End-User Perspective

Sudatta Chowdhury and Monica Landoni

Dept. of Computer and Information Sciences
University of Strathclyde
UK G1 1XH
{sudatta.chowdhury, monica.Landoni}@cis.strath.ac.uk

Abstract. Reports on what users experience when interacting with currently available news aggregator services. Five news aggregator services were chosen as the most representatives of emerging trends in this area of research and a combination of quantitative and qualitative methods were used for data collection involving users from the academic and research community. Forty-five responses were received for the online questionnaire survey, and 10 users were interviewed to elicit feedback . Criteria and measures for comparing usability of the chosen services were defined by the researchers based on the review of literature and a detailed study of the chosen news aggregator services. A number of desirable features of news aggregators were identified. Concluded that an ideal model could be designed by combining the usability features of TvEyes and the retrieval performance of GoogleNews.

Keywords: News aggregators, usability, evaluation, user study.

1 Introduction

In today's digital world, it is difficult to choose the right channel that can provide the required information in the desirable form with less effort and in a reasonable time. Web content aggregator services – individuals or organizations that gather web content, and/or applications, from different online sources for reuse or resale – can help users find the required information from a variety of information channels without much effort and time. There are two kinds of web aggregators: (1) those that simply gather material from various sources and put it on their Web sites, and (2) those that gather and distribute content, after doing the appropriate organization and processing, to suit their customers' needs [1]. A news aggregator usually collects news information from a variety of channels and summarizes in a pre-/user defined format. There is a need for a service that can provide seamless access to a variety of multimedia and multilingual digital news information resources with appropriate personalization, summarization, cross-lingual and other facilities. Keeping this requirement in mind, an EU funded project, called REVEAL THIS [2], has been undertaken by a multidisciplinary research team. As part of the project, a

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 481 – 484, 2006.

survey was conducted to know what users expect from a news aggregator service; what users get after using a chosen service and what they comment on the available useful features. This paper reports on a study that aimed at finding out the essential and most usable/appealing features of news aggregator services as perceived by the end-users.

2 Methodology

Several researchers have defined usability of digital libraries with emphasis on users and context (see for example [3,4,5]). Recent studies on usability testing with specific references to digital libraries include those of Allen [6], Blandford and Buchanan [7], Blandford [8], Dickstein and Mills [9], and Mitchell [10]. These papers have helped in designing the criteria and measures for comparing usability of the chosen news aggregator services. Eighteen usability criteria were identified for this study, and the users were asked to comment on each of them for the chosen five news aggregator services, i.e., Headlinespot (http://www.headlinespot.com/), Tveyes (http://www.tveyes.com/), Newsburst (http://www.newsburst.com), Google News (http://news.google.com/), and Awasu (http://www.awasu.com/). The survey used an online questionnaire and personal interviews to collect data involving university postgraduate students, researchers and academics.

3 Findings of the User Study

Results were gathered after analyzing all 45 responses from the online questionnaire survey, and ten in-depth interviews. Questions were asked about usability of the system, and the users were given a scale (1-10) to mark the various service features. Table 1 shows the users' comments on the services.

4 Desired Features of a 'Dream' News Aggregator Service

This survey provides some interesting findings with regard to the features of the selected news aggregator services, and these may provide useful guidelines for improving the existing news aggregator services. The results of this survey indicate that among the services evaluated TvEyes had the best reviews from users in all areas but engagement and performance at retrieval stage followed by Google News. The Headlinespot service was criticized across relevance, satisfaction and efficiency, ranking it behind TvEyes and GoogleNews. These suggest that TvEyes and GoogleNews are the two best services complimenting each other in terms of usability and functionalities. From this survey it emerges that a service with the usability of TvEyes and the retrieval performance of Google News could be a 'dream' news aggregator service. A list of the features that are desired by the end-user for the 'dream' news aggregator service are as follows:

- easy to use, fast, engaging, helpful and interesting
- good look and feel
- clear and expressive help
- quality information
- advanced search facilities
- better presentation
- avoid repetition of retrieved information
- show time of last update
- personalisation features
- regional coverage
- alert service
- descriptive subject categories
- results for search terms categorised by content
- easy to find subjects among existing categories
- easy to find information using the tool provided
- easy to work with the retrieved material
- make users feel confident that all the relevant information are found

Table 1. User responses to the features of the services (average scores in the 10-point scale)

Features	**Google News**	**Headline post**	**TVEyes**	**AWASU**	**News Burst**
Ease of use	6.6	6.3	7.8	7.75	6.3
Engaging	4.7	5.5	7.5	5.75	5
Frustrating	6	4	4.8	4.75	5
Helpful	6	6.1	7	6.25	5.5
Interesting	5.6	6	7.25	7.25	5.3
Likable	5.5	7.5	5.8	5.5	5
Useful	6.1	6.46	6.8	7.25	5.8
Annoying	4.4	4.2	3.1	5	5
Unpleasant	4.4	3.4	2.5	4.5	3.8
Descriptive broad categories	7.5	5.4	7	6.5	3.3
Subjects are easy to find in the categories	5.7	5.25	6	6.5	5.6
Hard to concentrate on searching infn. due to distractions provided by the service/tool	4.4	5	5.3	5.75	5
Confusing to work with the tool	5.4	3.6	3.75	5	5.1
Easy to find specific information	5	4.8	5.75	5.25	4.6
Frustrating to work with the provided tool	5.7	4.8	3.25	7.25	6.4
Tasks can be accomplished quickly	6.3	5.9	6.5	5.75	5.8
Easy to work with retrieved material	6.5	5.7	6.6	6	6.5
Confident that the search was exhaustive	6	4.9	4.8	5	4.8

Acknowledgements

This research is part of an EU funded research project called REVEALTHIS (Project no. FP6-IST-511689). Thanks are due to the partners of the REVEAL THIS project for providing valuable comments on the design of the survey, and special thanks are due to the survey participants.

References

1. Content aggregator – a Whatis.com definition. Available http://searchwebservices.techtarget.com/sDefinition/ 0,,sid26_gci815047,00.html
2. REVEAL THIS (FP6-IST-511689).Retrieval of Video and Language for the Home User in an Information Society. Sixth Framework Programme Proposal /Contract no. Unpublished
3. Borgman, C. and Rasmussen, E. (2005) Usability of digital libraries in a multicultural environment. In: Theng, Ying-Leng and Foo, Schubert. Eds. Design and usability of digital libraries: case studies in the Asia-Pacific. London: Information Science Publishing, 270-84.
4. Chowdhury, G.G. (2004) Access and usability issues of scholarly electronic publications. In: Gorman, G.E. and Rowland, F. eds. Scholarly publishing in an electronic era. International yearbook of Library and Information management, 2004/2005. London: Facet Publishing, 77-98.
5. Dillon, A. (1994). Designing usable electronic text: ergonomic aspects of human information usage. Bristol: Taylor and Francis.
6. Allen, M. (2002) a case study of the usability testing of the University of South Florida's virtual library interface design. Online Information Review, 26(1), 40-53.
7. Blandford, A. and Buchanan, G. (2003) Usability of digital libraries: A source of creative tensions with technical developments. TCDL Bulletin. Available: http://www.ieee-tcdl.org/Bulletin/current/blandford/blandford.html
8. Blandford, A. (2004) Understanding user's experiences: evaluation of digital libraries. Presented at the DELOS workshop on evaluation of digital libraries Padova, Italy. Available: http://www.delos.info/eventlist/wp7_ws_2004/Blandford.pdf
9. Dickstein, R. and Mills, V. (2000) Usability testing at the University of Arizona Library: How to let the users in on the design. Information Technology and Libraries, 19(3), 144-51.
10. Mitchell, S. (1999) Interface design considerations in libraries. In: Stern, D. ed. Digital libraries: philosophies, technical design considerations, and example scenarios. New York: The Haworth Press, 131–81.

DIAS: The Digital Image Archiving System of NDAP Taiwan

Hsin-Yu Chen, Hsiang-An Wang, and Ku-Lun Huang

Institute of Information Science, Academia Sinica, Taipei, 115, Taiwan
{kwakwai8, sawang, kulun}@iis.sinica.edu.tw

Abstract. The Digital Image Archiving System (DIAS) was developed by the National Digital Archives Program, Taiwan. Its major purpose is to manage and preserve digital images of cultural artifacts and provide the images to external

DIAS uses the DjVu image technique to solve the speed and distortion problems that arise when browsing very large images on the Internet. It also provides an online, real-time visible watermark appending function for digital image copyright protection, and uses image copy detection techniques to track illegal duplication.

Currently, DIAS manages a vast number of digital images and can be integrated with metadata archiving systems to manage digital images and metadata as a complete digital archiving system. We are developing digital image data exchange, heterogeneous system integration, automatic image classification, and multimedia processing technologies to improve DIAS.

Keywords: Copy detection, digital archive, digital image, DjVu, watermark.

Article

The Digital Image Archiving System (DIAS) [3] is the image management system of the National Digital Archives Program, Taiwan [7]. Its major objectives are to preserve valuable digital images and relevant information about the images, and provide the content to external metadata archiving systems on request. DIAS can be integrated with metadata archiving systems to build a complete digital archiving platform, as shown in Figure 1.

Although metadata archiving systems can handle texts efficiently, the images are in many different formats. In addition, the files are large, computation costs are high, a large storage space is required, and the files are difficult to analyze and record. Therefore, we designed DIAS as a specialized digital image management system in which the hardware has better processing ability, and there are sophisticated digital image processing techniques for managing digital image content. There is also more storage space.

DIAS provides basic control functions for processing digital images, including uploading, downloading, file content management, moving, deleting, spinning, image description, browsing, searching, and user/system authentication. In addition, because digital images of antiques are often large, transferring them over the Internet is not feasible

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 485 – 487, 2006.

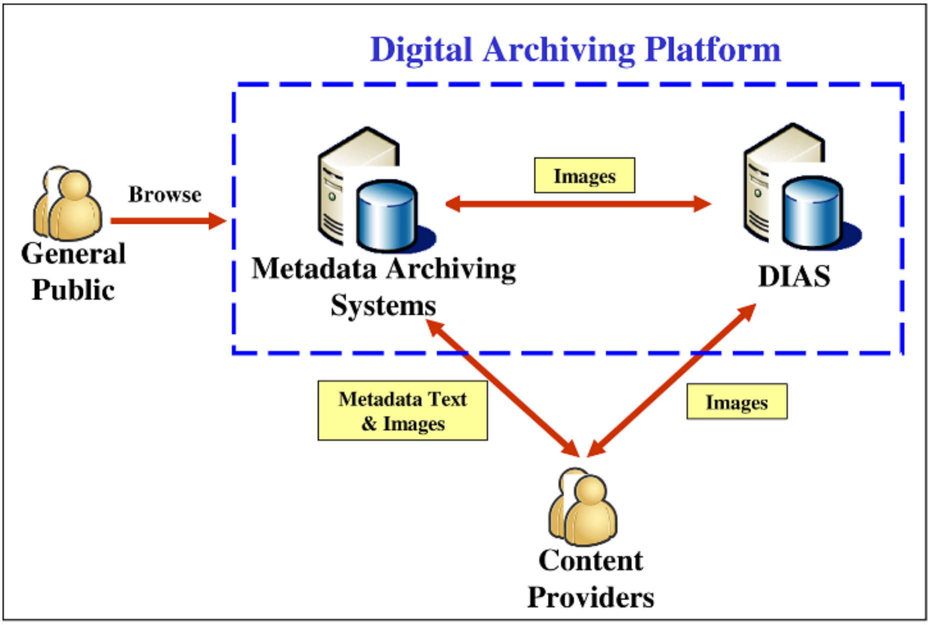


Fig. 1. The integration of DIAS with metadata archiving systems

Therefore, conversion of large files into smaller files is necessary so that users can browse them on the Internet. For this reason, we evaluated many digital image formats and decided to utilize the DjVu format in DIAS. DjVu uses a wavelet-based compression technique [2] to reduce digital images effectively. The resulting images can then be enlarged by the user without losing their original detail. We therefore implemented the function to convert original digital image files into DjVu images files for easy online use.

Since it is likely that digital images in the network environment will be duplicated illegally, a copyright protection mechanism is needed. We evaluated several DRM (Digital Rights Management) techniques and implemented some of them to protect the copyright of images [4]. In addition to user/system authentication, we developed a technique of adding visible watermarks in real-time online to indicate ownership without destroying the original images. We also experimented with an image copy detection technique [5] that compares image files rapidly to determine whether they have similar features, i.e., to determine if the original image has been copied. This enables owners to track illegal duplication automatically. We also use a wrapper-based DRE technique [6] to protect the digital rights. When a user downloads digital content from the network and views it on a player (e.g., a browser), the wrapper automatically monitors the user’s behavior. If the rules are violated, or the user refuses to be monitored by the wrapper, the content is rendered unavailable.

Currently, communication between DIAS and metadata archiving systems is through URLs (Uniform Resource Locators) and WebFTP (Web interface over File Transfer protocol) mechanisms on the Internet. When users wish to view images in the metadata archiving system, the system automatically requests the image data from

DIAS via the URL and presents the images to the user. Furthermore, if users choose to add new images to the metadata archiving system, the latter automatically transfers the images to DIAS via WebFTP and DIAS processes and stores them in the system. These mechanisms facilitate automatic communication between DIAS and various metadata archiving systems.

DIAS provides digital image management, a large image format conversion module, and DRM protection. It can be integrated with metadata archiving systems to reduce the burden of managing large amounts of archived image data. So far, DIAS has been implemented in ten metadata archiving systems of NDAP, and over 510,000 digital images have been preserved; the amount continues to increase.

DIAS uses a hierarchical classification structure to manage and show the vast number of images related to different topics. For easier management, searching, and browsing, DIAS allows users to customize their catalogues according to the attributes and hierarchical relationship of the images. Each image is stored in a classified catalogue. In order to manage and protect image content, the system allows different users to have different access rights to catalogues and images. The design of classified catalogues simplifies the management of the images, and speeds up searching and browsing.

The efficiency of DIAS could be affected by the volume of images loaded, the increasing number of users, and connections to many other metadata archiving systems. To enhance the system's scalability, in the future, we will use Web Services techniques to implement the distributed DIAS system, and exchange large amounts of image content and metadata with other archiving systems. In addition, we are designing an automatic image classification function for image searching and classification [1]. Finally, we will expand DIAS' multimedia processing capabilities into a comprehensive multimedia archiving system.

References

1. C. H. Li, C. Y. Chiu, and H. A. Wang, "Image Classification for Digital Archive Management", Proc. of 8th International Conference on Asian Digital Libraries (ICADL), pp. 81-89, Bangkok, Thailand, Dec. 2005.
2. DjVu Zone, "What is DjVu", http://www.djvuzone.org/wid/index.html
3. Digital Image Archiving System (DIAS), http://ndmmc2.iis.sinica.edu.tw/
4. H. Y. Chen, C. H. Li, C. H. Chiu, and W. L. Lin, "Implementation of Digital Rights Management with Digital Archiving System as An Example", The fourth workshop on Digital Archives Technology, pp. 93-100, Taipei, Taiwan, Sep. 2005.
5. J. H. Hsiao, C. S. Chen, L. F. Chien, and M. S. Chen, "Image Copy Detection via Grouping in Feature Space Based on Virtual Prior Attacks," International Conference on Image Processing, Atlanta, GA, USA, 2006.
6. J. H. Hsiao, J. H. Wang, M. S. Chen, C. S. Chen and L. F. Chien, "Constructing a Wrapper-Based DRM System for Digital Content Protection in Digital Libraries," Proceedings of 8th International Conference on Asian Digital Libraries, ICADL 2005, Bangkok, Thailand, pp. 375-379, December 12-15, 2005.
7. National Digital Archives Program (NDAP), Taiwan, http://www.ndap.org.tw/index_en.php

Distributed Digital Libraries Platform in the PIONIER Network

Cezary Mazurek, Tomasz Parkoła, and Marcin Werla

Poznan Supercomputing and Networking Center
Noskowskiego 10, 61-704 Poznań, Poland
{mazurek, tparkola, mwerla}@man.poznan.pl

Abstract. One of the main focus areas of the PIONIER: Polish Optical Internet program was the development and verification of pilot services and applications for the information society. It was necessary to create a base for new developments in science, education, health care, natural environment, government and local administration, industry and services. Examples of such services are digital libraries, allowing to create multiple content and metadata repositories which can be used as a basis for the creation of sophisticated content-based services. In this paper we are presenting the current state of digital library services in the PIONIER network, we shortly describe dLibra - a digital library framework which is the software platform for the majority of PIONIER digital libraries. We also introduce two content-based services enabled on PIONIER digital libraries: distributed metadata harvesting and searching and virtual dynamic collections.

Keywords: Digital libraries, service-oriented architecture, metadata harvesting, virtual collections, distributed searching, distributed resources syndication.

1 Introduction

The first digital library deployed in the PIONIER network was the Digital Library of the Wielkopolska Region (http://www.wbc.poznan.pl/) started in October 2002, as the result of the cooperation between Poznan Supercomputing and Networking Center (PSNC) and Poznan Foundation of Scientific Libraries. This library was also the first digital library based on the PSNC's dLibra Digital Library Framework. Currently in the PIONIER network there are 5 regional and 5 institutional dLibra-based digital libraries and at least three more will be available before the end of 2006[1].

All these libraries together create a platform of distributed digital libraries with over 30 000 digital objects (writing relicts, cultural heritage, regional and educational materials) and an average number of 1 000 concurrent users each moment. Creation of such platform is possible because of the service-oriented architecture of the dLibra framework. The framework functionality and architecture is briefly described in the

[1] Full list of dLibra-based digital libraries can be found on the dLibra project homepage (http://dlibra.psnc.pl/).

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 488 – 491, 2006.

next section. Digital libraries in the PIONIER network are also the basis for advanced content services enabled for educational and scientific users. At the end of the paper we describe two examples of such services: distributed metadata search and virtual dynamic collections.

2 dLibra Digital Library Framework Overview

The dLibra Digital Library Framework is the first Polish digital library software platform developed by Poznan Supercomputing and Networking Center (PSNC) as a part of the PIONIER programme (http://www.pionier.gov.pl/). The dLibra project was started in 1999, as a part of research in the field of digital libraries started in PSNC in 1996. The developed platform is currently the most popular digital library framework in Poland.

The dLibra Framework was designed to be a highly configurable software basis for digital libraries. dLibra-based digital library can be used to preserve, manage and access digital objects consisting of the content (text, sound, video, etc.) and the metadata in a user-defined schema. Digital objects in the dLibra Framework are organized in a hierarchical way using so called directories and group publications. In addition to present publications for the needs of Internet users and a distributed search mechanism, dLibra uses collections of objects defined by the digital library administrator.

A single dLibra instance has a multitier architecture presented below (Fig. 1).

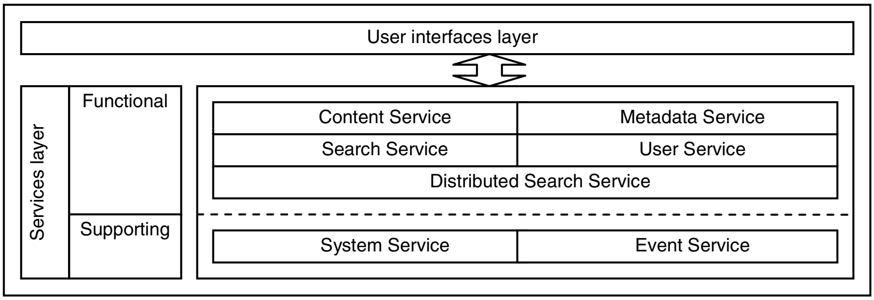


Fig. 1. Architecture of the dLibra framework with user interface layer and services layer

The main component of this architecture is a dLibra server separated into five functional level services (Metadata, Content, User, Search and Distributed Search Services) and two supporting level services (Event and System Services). Functional level services are responsible for the entire dLibra-based digital library functionality. Supporting level services were created to maintain inter-service mechanisms such as service resolving, authentication and communication.

Two client applications (User interfaces layer on Fig. 1) have access to the dLibra system. The first one is the Editor's/Administrator's Application. It allows users to add new objects to the library and manage all the gathered content. It also gives administrators a possibility to manage some core DL parameters like used metadata scheme, metadata dictionaries etc. The second client application is the web-based Reader's Application. It gives users read-only access to the library with such functionality as browsing, searching, news feeds etc. More detailed information about the dLibra architecture and functionality can be found in [1,2,3].

3 Content-Based Services in PIONIER Digital Libraries

The latest functionality of the dLibra Digital Library Framework provided for all PIONIER digital libraries included periodic metadata synchronization in the entire platform based on the OAI-PMH protocol [4]. This process is based on selective harvesting and information about the deleted records. Custom collections defined in digital libraries are transformed into OAI-PMH sets and custom metadata schemas are transformed to DCMES attributes. An additional feature required introducing platform-level metadata synchronization was a system of unique digital object identifiers, based on the OAI-identifier syntax. The functionality applied to all PIONIER digital libraries transformed those libraries into one distributed platform where each digital library has full information about all metadata in the platform. This metadata can be used by each digital library to provide its users with means to access all resources stored in all other digital libraries and to create new advanced content- and metadata- based services. Two of them are described below.

3.1 Distributed Metadata Search

Distributed metadata search is based on a periodic metadata synchronization process performed between all digital libraries in the platform. Each digital library indexes metadata harvested from remote libraries and allows users to search through the indexed metadata. The entire process is performed transparently to the end user and, as a result, the user receives one list of search results with both local and remote resources which matched the user query. Such distributed searching applications might be easily accessed by other network services or information portals also from the outside of the digital library framework. This approach formulated a new kind of content- and context-based service which might be used to create advanced network applications. Currently we are planning to extend search possibilities to include content-based content search. This will be based on a combination of OAI-PMH and MPEG-21 DIDL [5] and it is a subject of current R&D activities in PSNC.

3.2 Virtual Dynamic Collections

Virtual dynamic collections in the PIONIER digital libraries platform are basically collections containing elements from different digital libraries across the platform. Those collections are defined by users as conditions that should be met by digital objects metadata. When a new digital object is published in the platform, it propagates through all digital libraries and all defined virtual collections. Such mechanism does not require any additional work of either the digital library administrator or the editor, in contradistinction to static collections where all resources must by explicitly assigned to a collection by the digital library editor. The virtual dynamic collections mechanism is based on RSS feeds, allowing to access it from various user-selected external applications, portals and services. Such approach significantly increases the visibility of digital objects in the PIONIER network.

4 Conclusions

In this paper we have shortly presented the current state of distributed digital library services in the PIONIER network. The latest functionality provided to that services was metadata synchronization based on the metadata harvesting with the OAI-PMH protocol. It transformed the PIONIER digital libraries into a coherent distributed platform where each digital library became an access point to all resources stored in the PIONIER digital libraries. The implementation of the metadata synchronization between different instances of digital libraries is the basis for the development of new content-based services. We have presented two examples of such services enabled through the PIONIER platform for distributed digital libraries, provided for research and education users. By the end of 2006 we aim to develop further similar services. The next step that will allow for the creation of even more sophisticated services, will be a possibility to search through content gathered in the PIONIER digital libraries. This will be achieved by combining the OAI-PMH protocol together with the MPEG-21 DIDL standard for the exchange of the digital content and content-related metadata. Another group of complementary services also covers information services provided by Grid environments [6] and development of information grids.

References

1. Mazurek, C., Werla, M.: Distributed Services Architecture in dLibra Digital Library Framework. 8th DELOS Workshop on Future Digital Library Management Systems, Workshop Proceedings (2005)
2. Heliński, M., Mazurek, C., Werla M.: Distributed Digital Library Architecture for Business Solutions. EUROMEDIA'2005 Conference, Conference Proceedings (2005)
3. Mazurek, C., Werla, M.: Digital Object Lifecycle in dLibra Digital Library Framework. 9th DELOS Workshop on Interoperability and Common Service, Workshop Proceedings (2005)
4. Mazurek, C., Stroiński, M., Werla, M., Węglarz, J.: Metadata harvesting in regional digital libraries in PIONIER Network. TNC 2006 Conference, Conference Proceedings (2006)
5. Bekaert, J. et al.: Using MPEG-21 DIDL to Represent Complex Digital Objects in the Los Alamos National Laboratory Digital Library. D-Lib Magazine Vol. 9 No 11 (2003)
6. Kosiedowski, M.; Mazurek, C; Werla, M.: Digital Library Grid Scenarios. European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology, Workshop Proceedings (2004)

EtanaCMV: A Visual Browsing Interface for ETANA-DL Based on Coordinated Multiple Views

Johnny L. Sam-Rajkumar, Rao Shen, Naga Srinivas Vemuri,
Weiguo Fan, and Edward A. Fox

Digital Library Research Laboratory, Virginia Tech, USA
{johnny, rshen, nvemuri, wfan, fox}@vt.edu

Abstract. Visual interfaces for digital libraries (DLs) provide support for insightful browsing, presentation of search results in a browsing platform, and analysis of records in the DL. We propose the demonstration of a visual interface to ETANA–DL, a growing union archaeological DL. Our interface EtanaCMV is based on a uniform multiple view design and facilitates browsing of DL records that are multidimensional, hierarchical, and categorical in nature. We use distinct panels to allow flexible browsing across multiple dimensions. Bars in each panel denote the various categories in each dimension. EtanaCMV will give the users a quick overview of the collections in the DL during browsing in addition to showing relationships in the harvested collections. Coordination between multiple views is used to present more insights into the data.

1 Introduction

Records in a digital library may be browsed by several dimensions. For example, archaeological records may be browsed by the "object type" dimension or the "time" dimension. The same set of records may be categorized in different ways according to the different dimensions. Textual browsing systems are easy to navigate, but they greatly limit the speed of gaining insight about the available records. Further, it often is hard to present clearly the relationships between categories using textual interfaces.

Visual interfaces for digital libraries:

- Present more insight into the distribution of records in each category within a dimension and across all dimensions. Accordingly, Citiviz [1] and EtanaViz [2] interfaces are based on multiple views, while Envision [3] and GriDL [4] employ a 2-D scatter plot like design. Relational Browser++ [5] has uniform multiple views with bars to denote categories.
- Enable graphical display of search results over the browsing interface.
- Support visually-enhanced analysis of collections in the DL.

We present the initial design of our visual interface (see Fig. 1) and conclude with a description of ongoing work on this interface to support more DL services.

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 492 – 494, 2006.

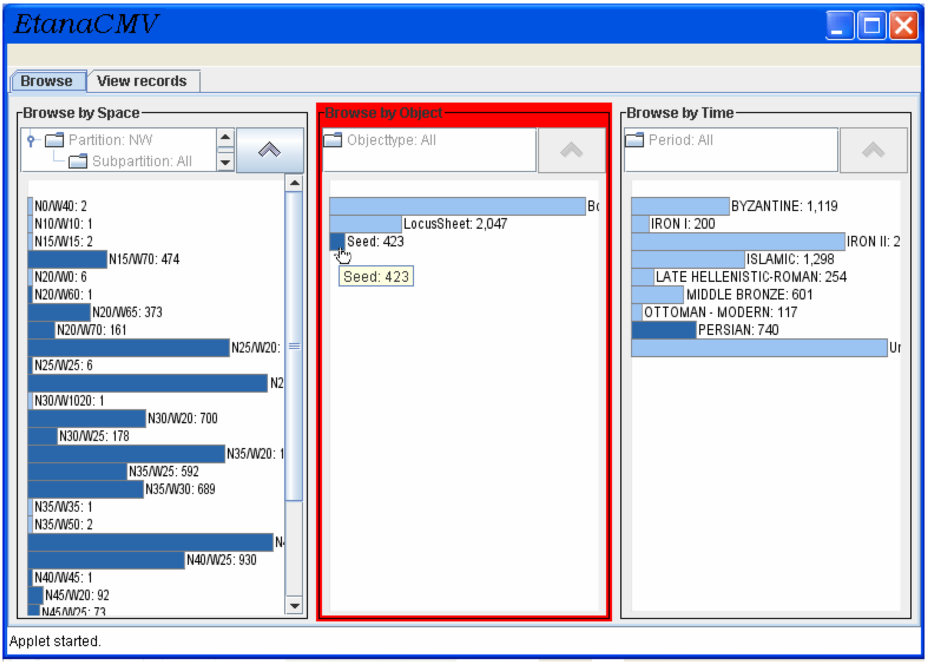


Fig. 1. Initial Design of EtanaCMV

2 Design

The archaeological records from ETANA-DL [6] can be browsed by three dimensions: Space, Object type and Time. Each dimension is hierarchical. For example, the space dimension fits the hierarchy of site, partition, sub-partition, locus, and container.

There is one panel for each dimension. The bars show the categories in each dimension and the length of the bars indicate the number of records in each of these categories. Moving the mouse over a panel causes that panel to be highlighted, indicating the current browsing dimension. The panels are scrollable depending on the number of bars in them. The three views are coordinated through the visualization technique called "brushing and linking". For example, moving the mouse over the bar representing "Seed" category under "Object type" dimension causes the "Persian" bar to highlight under the "Time" dimension. This is because records of seed object type are found in the Persian era. Users can click on a bar to drill down into the hierarchy in a particular dimension. The tree view on the top of each chart shows the navigation path of the user in the current dimension. The button next to the tree view allows users to navigate to higher hierarchical levels. Users will be able to view the records within the current browsing context by clicking the "View records" tab.

3 Conclusion and Future Work

We believe that this design will give users a flexible and insightful visual browsing platform for navigating through the hierarchies and across dimensions without getting lost. The support for hierarchical browsing distinguishes our interface from Relational Browser++. Before the conference we will conduct a formative evaluation of the system to refine the interface. We also plan to extend this interface to coordinate with the search service.

Acknowledgements. This work is funded in part by the National Science Foundation (ITR-0325579).

References

1. Kampanya, N., Shen, R., Kin, S., North, C., Fox, A.: CitiViz: A Visual User Interface to the Citidel System, Lecture Notes in Computer Science v. 3232, Jan 2004, pages 122 - 133
2. Shen, R., Vemuri, N. S., Vijayaraghavan, V., Fan, W., and Fox, E. A.: EtanaViz: A Visual User Interface to Archaeological Digital Libraries, Technical Report TR-05-14, Computer Science, Virginia Tech (2005)
3. Shneiderman, B., Feldman, D., Roseand, A., and Grau, X. F.: Visualizing digital library search results with categorical and hierarchical axes, Proceedings of the 5th International Conference on Digital Libraries (DL '00), ACM Press, New York, NY, 2000, pages 57-66
4. Heath, L. S., Hix, D., Nowell, L.T., Wake, W.C., Averboch, G.A., Labow, E., Guyer, S.A., Brueni, D.J., France, R.K., Dalal, L., and Fox, E.A.: Envision: A user-centered database of computer science literature. Communications of the ACM, 38(4):52--53, April 1995.
5. Zhang, J., and Marchionini, G.: Evaluation and Evolution of a Browse and Search Interface: Relation Browser++, Proceedings of the 2005 National Conference on Digital Government Research, DG.O 2005, Atlanta, Georgia, USA, May 15-18, 2005, pages 179-188
6. U. Ravindranathan. Prototyping Digital Libraries Handling Heterogeneous Data Sources - An ETANA-DL Case Study. Masters Thesis. Computer Science, Virginia Tech, Blacksburg VA, April 2004, http://scholar.lib.vt.edu/theses/available/etd-04262004-153555/

Intelligent Bibliography Creation and Markup for Authors: A Step Towards Interoperable Digital Libraries

Bettina Berendt, Kai Dingel, and Christoph Hanser

Institute of Information Systems, Humboldt University Berlin,
D-10178 Berlin, Germany
http://www.wiwi.hu-berlin.de/~berendt

Abstract. The move towards integrated international Digital Libraries offers the opportunity of creating comprehensive data on citation networks. These data are not only invaluable pointers to related research, but also the basis for evaluations such as impact factors, and the foundation of smart search engines. However, creating correct citation-network data remains a hard problem, and data are often incomplete and noisy. The only viable solution appear to be systems that help authors create correct, complete, and annotated bibliographies, thus enabling autonomous citation indexing to create correct and complete citation networks. In this paper, we describe a general system architecture and two concrete components for supporting authors in this task. The system takes the author from literature search through domain-model creation and bibliography construction, to the semantic markup of bibliographic metadata. The system rests on a modular and extensible architecture: VBA Macros that integrate seamlessly into the user's familiar working environment, the use of existing databases and information-retrieval tools, and a Web Service layer that connects them.[1]

Keywords: User interfaces for Digital Libraries, Collection building, management and integration, System architectures, integration and interoperability.

The importance of citation networks for science (for a recent overview, see [4]) is reflected in a large number of non-commercial and commercial services that archive various combinations of bibliographic metadata, metadata on citations, and full texts. These are beginning to interlink their data, and many of them are key drivers of Open Access (e.g., www.informatik.uni-trier.de/~ley/db, repec.org; scientific.thomson.com/products/{sci|scci}, portal.acm.org/guide.cfm; scholar.google.com; www.arxiv.org; citeseer.ist.psu.edu, www.citebase.org, www.slac.stanford.edu/spires/hep/, portal.acm.org/dl.cfm).

A major problem is that comprehensive manual markup is too costly (and subject to human error), while autonomous citation parsing and indexing (see

[1] For an extended version, see www.wiwi.hu-berlin.de/~berendt/DL

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 495–499, 2006.

[5] and the methods used by the other services named above) is limited by the visibility of documents, the heterogeneity of citation styles, and the recognition rates of parsing algorithms. Errors propagate, impede literature search and domain understanding, and may damage author standing.

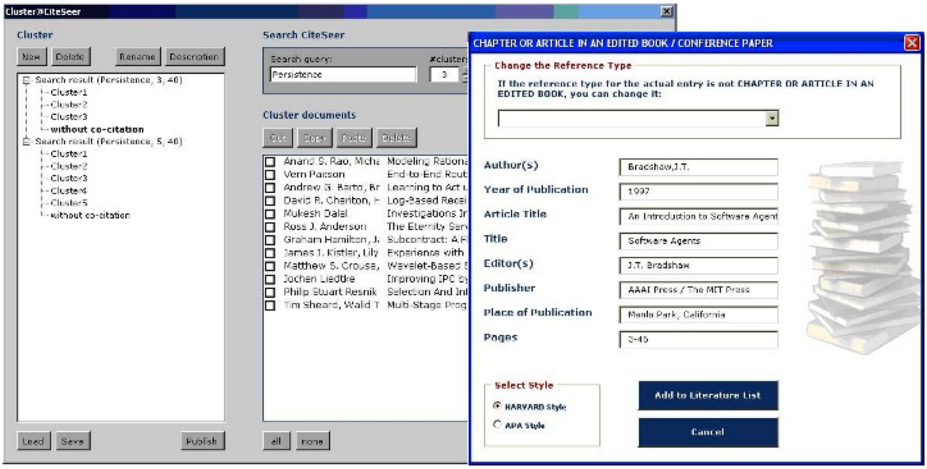
Only authors can and must create correct high-quality (meta)data. To obviate citation parsing and its errors, authors should ideally supply structured metadata (e.g., in BibTex or EndNote format, or by using templates such as those available at `edoc.hu-berlin.de/e_autoren`). A persistent, globally unique identifier such as a URN or DOI would be even better, but at present this remains an elusive goal in the heterogeneity of the Web.

Our prior research [2,1] has shown that structured-bibliography creation is little known, unpopular, and/or performed inconsistently. Therefore, we propose to (a) support authors in their familiar environments and writing styles, (b) employ machine intelligence and interactivity to improve quality, and (c) motivate authors to invest the remaining additional effort by showing them what they can gain from citation metadata. To the best of our knowledge, our system is unique in this integration of existing methods (bibliometric analyses, information extraction, and interface design) and process-comprehensive author support.

Requirements and system architecture. An intelligent author-support system should support the main elements and processes of scientific writing. Literature search, bibliography maintenance, domain understanding by domain structuring as well as the creation of one's own bibliography are key elements of scientific writing. (For further details and additional system components, see `http://www.wiwi.hu-berlin.de/~berendt/DL`.) The system should integrate itself seamlessly into the user's everyday working environment, and it should be modular, easily maintainable, and extensible. It should offer access to the huge and distributed literature databases available online.

As our studies (and observations elsewhere) show, the vast majority of authors use Microsoft Word for producing texts, work on low-end to medium PCs, and want to avoid installation activities. Therefore, we employ a combination of VBA Macros, Web Services, and Web-independent backbone intelligence. Thus, only the GUIs shown in this paper are MS-Word-specific; the intelligent services reached via Web Services can also be called from other interfaces.

Literature search and domain structuring. The screenshot on the next page (left side) illustrates the basic functionalites: Given a search term, a bibliographic database is searched, matching items are returned, and they are clustered. Each source is associated with a hyperlink to the full text from the online database. The user is encouraged to label (rename) the clusters and to modify the grouping to both reflect and develop his perception of the domain in terms of a topic structure (cut, copy, paste, delete). He can include the results in personal documents, provide an additional description, and publish the results to make them available also to others. Publication results are represented in XML to retain as much semantic structure as possible, and visualized as HTML to maximize visibility and accessibility. The hyperlinks to the online database are a popular proxy for the



Search result (Persistence, 3, 40)
- Cluster1
- Cluster2
- Cluster3
- **without co-citation**

Search result (Persistence, 5, 40)
- Cluster1
- Cluster2
- Cluster3
- Cluster4
- Cluster5
- without co-citation

Search query: Persistence
#clusters: 3

□	Anand S. Rao, Micha	Modeling Rationa
□	Vern Paxson	End-to-End Rout
□	Andrew G. Barto, Br	Learning to Act u
□	David R. Cheriton, H	Log-Based Recei
□	Mukesh Dalal	Investigations Ir
□	Ross J. Anderson	The Eternity Serv
□	Graham Hamilton, J.	Subcontract: A F
□	James J. Kistler, Lily	Experience with
□	Matthew S. Crouse,	Wavelet-Based S
□	Jochen Liedtke	Improving IPC by
□	Philip Stuart Resnik	Selection And Inf
□	Tim Sheard, Walid T	Multi-Stage Prog

If the reference type for the actual entry is not CHAPTER OR ARTICLE IN AN EDITED BOOK, you can change it:

Author(s)	Bradshaw,J.T.
Year of Publication	1997
Article Title	An Introduction to Software Agent
Title	Software Agents
Editor(s)	J.T. Bradshaw
Publisher	AAAI Press / The MIT Press
Place of Publication	Menlo Park, California
Pages	3-46

● HARVARD Style
○ APA Style

elusive publication URNs; this supports the re-use of existing citation networks further. Users can save and re-load results for further processing.

We use the CiteSeer database because of its broad coverage[2] and rich structure, and because it offers an OAI interface.

Clustering is based on co-citation as a long-validated and popular indicator of similarity and domain development [6]. CiteSeer offers a localised co-citation search that starts from a given document and returns those documents that are co-cited with it. Our system extends this by a more global view (the context of the immediately-relevant documents), a context-aware similarity measure (the Jaccard coefficient rather than the absolute number of co-citations), and the support of domain-model construction.

The user interface is a VBA macro that interacts with a php Web service, which accesses further information sources: (a) The search term is transformed into an HTTP request to CiteSeer; the result set D's document IDs are extracted by a wrapper. (b) All documents that cite any document from D are retrieved from a local mirror of the CiteSeer database, and the co-citation matrix is compiled. (c) Up-to-date metadata of D for result presentation (author, title, CiteSeer URL, etc.) are retrieved via CiteSeer's OAI interface. (d) The documents in D are clustered using CLUTO (`www.cs.umn.edu/~karypis/cluto`) with hierarchical single-linkage clustering [8], the number of clusters set to *min(user-defined number,* $|D| - 1$*)*, and the Jaccard coefficient. This similarity measure was first used in co-citation analysis by [7]; among other advantages, it precludes non-citing documents from inducing similarity. If present, isolated documents [8] are put into an additional cluster called "without co-citation" to avoid arbitrary assignments and to show all of the relevant literature on a topic [3].

[2] The use of further sources for integrating other disciplines and supporting federated search is the subject of future work.

Bibliography creation. To create metadata markup for a reference list, the author can mark the whole list with the mouse to receive a series of formatted bibliography entries as proposals (see screenshot on the right). Errors in automatic recognition can easily be spotted and corrected. When the user has accepted or corrected the system proposal, the macro writes a surface text into the Word document that is formatted according to the chosen citation style (here, Harvard or APA) and a metadata markup that contains the correct field entries. Our system creates DiML (`http://edoc.hu-berlin.de/diml/`), which allows the encoding of metadata for arbitrary sources (articles, books, but also datasets, ...); a generalisation to other markup schemes is straightforward.

The VBA macro calls a php Web service, which issues system calls to perl scripts of two programs: CiteSeer and ParaTools. Both are instances of template-based approaches, which are the currently dominant approach to autonomous citation indexing in large real-world repositories because they are scalable and do not need a training phase (for a recent survey of other approaches, see [9]).

Information extraction starts with the CiteSeer code. As an inspection of the CiteSeer Web site shows, the regular expressions used in this code are fairly effective at extracting author, year, and title information. At present the CiteSeer system does not extract further bibliographic information, probably because this information suffices for the tasks at hand: author/year/title are used to build the citation matrix, and the full text is used for keyword search. On the CiteSeer Web site, the extracted bibliographic information is shown as a sparse Bibtex entry; it is up to the community to add more information manually.

To fill missing slots, we use ParaTools (`http://paracite.eprints.org`), which are the basis of the software behind Citebase. In the ParaTools, alternative templates of a reference record, and alternative templates of each of its parts (author, title, etc.), are offered. Each of these regular-expression templates has two weights (reliability, concreteness). From all matching templates, the one with the highest weight combination is chosen. The template library can be extended.

Evaluation. In a random sample of 172 references from the automatically generated CiteSeer repository and from the hand-curated institutional repository `edoc.hu-berlin.de`, author/year/title information was identified correctly for 89% of the CiteSeer references and for 61% of the EDOC references (one problem is that author recognition is partly lexicon-based, which leads to difficulties with German names). Further bibliographic information was correctly identified by ParaTools for 17% of the sample. This result is likely to be a lower bound: for further tests, we plan to extend the template database shipped with ParaTools by typical citation styles found among our users. The bibliography creation component was motivated by earlier validations of the employed clustering methodology (see above); in addition, preliminary user tests suggest that subjective measures of utility and "correctness" will be more adequate for evaluating this component than objective measures of cluster quality.

Ultimately, success will rest on the users' satisfaction with the system. This requires large-scale user testing that we have recently begun.

Future work will exploit machine learning to identify publication similarity, to propose cluster labels, identify different instances and different versions of the same document, and to personalize the system.

References

1. Berendt, B. (2005). Understanding and Supporting Volunteer Contributors: The Case of Metadata and Document Servers. In *Proc. Knowledge Collection from Volunteer Contributors AAAI 2005 Symposium* (pp. 106-109). `www.wiwi.hu-berlin.de/~berendt/Papers/SS505BerendtB.pdf`
2. Berendt, B., Brenstein, E., Li, Y., & Wendland, B. (2003). Marketing for participation: How can Electronic Dissertation Services ... In *Proc. ETD 2003.* `edoc.hu-berlin.de/etd2003/berendt-bettina/`
3. Braam, R.R., Moed, H.F. & van Raan, A.F.J. (1991). Mapping of Science by Combined Co-Citation and Word Analysis. (I & II) *J. of the Amer. Soc. for Inform. Science, 42(4),* 233–266.
4. Chen, C. (2003). *Mapping Scientific Frontiers.* London: Springer.
5. Lawrence, S., Giles, C.L. & Bollacker, K.D. (1999). Digital Libraries and Autonomous Citation Indexing. *IEEE Computer, 32,* 67–71.
6. Small, H. (1973). Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents. *J. of the American Soc. for Inform. Science, 24(4),* 265–270.
7. Small, H. & Greenlee, E. (1980). Citation Context Analysis of a Co-citation Cluster: Recombinant-DNA. *Scientometrics, 2(4),* 277–301.
8. Small, H. & Griffith, B.C. (1974). The Structure of Scientific Literatures, I: Identifying and Graphing Specialities. *Science Studies, 4(1),* 17–40.
9. Yong, K.N. (2005). *Citation parsing using maximum entropy and repairs.* Dept. of Computer Science, Nat. Univ. of Singapore. `wing.comp.nus.edu.sg/publications/theses/yongKiatNgThesis.pdf`

Introducing Pergamos: A Fedora-Based DL System Utilizing Digital Object Prototypes

George Pyrounakis[1,2], Kostas Saidis[1], and Mara Nikolaidou[2],
and Vassilios Karakoidas[2]
forky@libadm.uoa.gr, saiko@di.uoa.gr, mara@di.uoa.gr,
bkarak@aueb.gr

[1] Department of Informatics and Telecommunications
[2] Libraries Computer Center
University of Athens
University Campus, Athens, 157 84, Greece

Abstract. This demonstration provides a "hands on" experience to the "internals" of Pergamos, the University of Athens DL System. Pergamos provides uniform high level DL services, such as collection management, web based cataloguing, browsing, batch ingestion and automatic content conversions that adapt to the underlying digital object type-specific specialities through the use of *Digital Object Prototypes* (DOPs). The demonstration points out the ability of DOPs to effectively model the heterogeneous and complex material of Pergamos. Special focus is given on the inexpensiveness of adding new collections and digital object types, highlighting how DOPs eliminate the need for custom implementation.

1 Introduction

Pergamos is the Digital Library System we developed for handling the heterogeneous and complex material of the University of Athens, originating from numerous sources, including the Senate Archive, the Theatrical Collection, the Folklore Collection and the Papyri Collection, to name a few. Pergamos is a web-based Digital Library implemented in Java that builds upon Fedora repository [2].

Pergamos provides a powerful digital object manipulation mechanism based on *Digital Object Prototypes* (DOPs) [1]. DOPs focus on the uniform resolution of digital object typing issues in an automated manner, releasing cataloguers, developers and collection designers, from dealing with the underlying typing complexity manually. All digital object typing information is expressed in terms of DOPs. The latter capture and express digital object typing requirements in a fine-grained manner, while they deploy a uniform "type conformance" implementation that makes all digital objects conform to their corresponding DOP specifications automatically. This way, the definition of new collections and respective digital object types is performed in a straightforward fashion, requiring no custom implementation or code development.

DOPs provide a detailed specification of: (a) the metadata sets used by the digital object type at hand (b) the digital content formats supported by this

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 500–503, 2006.

type, (c) the relationships in which instances of this type are allowed to participate and (d) the behaviors that all instances of this type should expose. DOPs are defined in terms of XML. The `DO Dictionary` depicted in Figure 1, loads the DOP XML definitions during DL startup. It then translates the DOP supplied definitions into Java artifacts that are exposed to higher level application logic through the DOPs API. All digital objects are associated with DOPs. At runtime, the `DO Dictionary` loads stored digital objects from the underlying repository and generates their corresponding digital object instances that automatically conform to the object's DOP. The details of the underlying repository remain hidden as all application logic's functionality is directed through digital object instances. The ability to expose "typefull" instances to the services of the application logic allows us to generate single, uniform service implementations which are capable to operate upon any DOP-defined type of material.

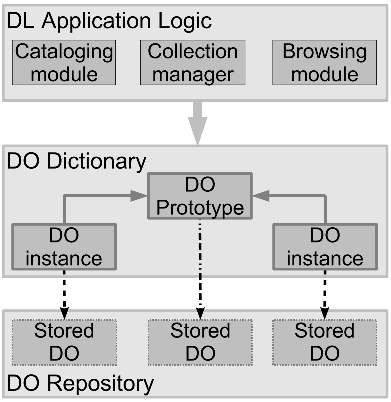


Fig. 1. Pergamos 3-tier architecture incorporating the "type enforcement" layer of DO Dictionary [1]

2 DOP-Based Pergamos Features

The demonstration consists of a "mixed" viewpoint approach on Pergamos web based DL services, elaborating on end-user's and cataloguer's, designer's and developer's perspectives. The demonstration pinpoints how the use of DOPs allows us to deal with important DL development issues in a uniform yet adaptive manner. We particularly emphasize on the following Pergamos features.

Collection Management and Complex Objects
For uniformity reasons, we treat collections as digital objects. DOPs support aggregation relationships – the objects of one type are allowed to "contain" objects of another type. A collection object is allowed to contain other collection objects, generating a collection hierarchy. The root of the hierarchy is the Digital Library itself, a "super collection" object containing all other collections.

DOPs are defined in the context of a specific collection. For example, Folklore Collection consists of instances of the `Notebook`, `Chapter` and `Page` DOPs. DOPs are supplied with fully qualified identifiers such as `folklore.Page` and `folklore.Notebook`, allowing us to support user-defined types of objects with collection-pertinent scopes. Thus, although the collection of the Senate Archive's Session Proceedings consists of objects belonging to the `Folder`, `Session` and `Page` DOPs, the latter is distinguished from the `folklore.Page` through having the `senate.Page` fully qualified identifier.

The addition of a new collection refers to specifying the individual DOPs that model each different type of material this new collection supports. Additionally, the DL designer is able to add sub-collection objects in the collection at hand, specifying each sub-collection's supported DOPs recursively. This way, a collection is made up of the digital object instances belonging to the DOPs the collection supports. An instance is either "added" to the collection explicitly or implicitly, through belonging to a DOP of one of the collection's sub-collections. DOPs also support complex objects in a same manner. Senate Archive's `Sessions` are modelled as complex objects that are allowed to contain `Page` objects.

Metadata Handling and Cataloguing Capabilities

DOPs specify the metadata used for each different digital object type in a fine-grained manner. Each digital object type may contain one or more metadata sets for descriptive or administrative purposes. Each metadata set specification in a DOP contains one or more metadata element definitions. For each metadata element, a DOP provides: its identifier and multi-lingual labels and descriptions along with additional element characteristics that assist in the proper treatment of its values at runtime.

The behavioral characteristics we support are:

- `isMandatory`, that directs instances to forbid null values for the element,
- `isRepeatable`, that directs instances to render the element values in a list,
- `defaultValue`, that directs instances to supply this value to the element if the cataloger has not explicitly provided another value
- `validation`, that executes the user-supplied validation plugin for enforcing desired constraints on the element's value.

DOPs also support the definition of mappings among elements of different metadata sets. For example, the archival nature of the Senate Archive's Session Proceedings is modelled as follows. We use the `dc` and `ead` metadata sets for `Folder` and `Session` objects. `dc` refers to a qualification of the DC elements, while `ead` follows the principles of EAD without encoding the Finding Aid in its entirety. Our ability to define and handle metadata sets and their respective elements in a type-specific manner enables us to generate a uniform implementation of the web-based Cataloguing service that effectively copes with all Pergamos material. The Cataloguing service can generate detailed metadata element representations for all types of objects in a unified way by exploiting the specifications residing in the object's DOP.

Automatic Content Conversions and Batch Ingestion

DOPs provide a detailed definition of the file formats supported by each different object type. For example, the `senate.Page` DOP specifies that its instances should consist of a high quality TIFF file held for preservation purposes, a lower quality JPEG file used for web display and a thumbnail JPEG image used for browsing.

We use digital content specifications of DOPs to automate content conversions. Each file specification in a DOP is defined as `primary` or `derivative`. `primary` file format specifications provide conversion information that is used by the respective instances to automatically convert the `primary` file format to its corresponding DOP-defined `derivatives`. For example, the `senate.Page` contains conversion specifications that allows its instances to automatically generate the JPEG images from the high quality TIFF image, whenever the latter is either ingested or replaced by the user.

Moreover, we use DOPs to generate effective batch content ingestion for diverse types of objects. We model `senate.Sessions` as containers of `senate.Page` objects. The `senate.Session` DOP provides a `container` file format specification that allows `Session` instances to automatically create `senate.Page` objects from a suitable user-supplied zip file. The batch ingestion process is invoked when the user uploads a zip archive to a `senate.Session` instance. If the archive contains files that belong to the `senate.Page primary` format (TIFF), the `senate.Session` instance automatically creates new `Page` objects for each TIFF file. Then it saves each TIFF file to its corresponding `Page` object, triggering the `senate.Page`'s automatic conversions described above.

Browsing and Searching

The hierarchical structure of digital material generated by the use of DOPs is reflected in Pergamos web-based browsing facility. Although the Browsing service resides in a uniform implementation, objects belonging to different types are displayed according to their corresponding type's requirements. Browsing service fetches the `browseView` behavior on each instance and the latter interprets the call in a DOP-defined manner automatically.

Pergamos search capabilities reflect the ones provided by FEDORA. However, the use of DOPs allows us to provide additional search functionality to our end users, allowing them to limit search results on selected collections, sub-collections or types. Moreover, we use FEDORA's built-in DC-based searching to support cross-collection searches, yet we are able to provide enriched metadata to our end users by exploiting the mappings capabilities of DOPs.

References

1. K. Saidis, G. Pyrounakis, and M. Nikolaidou. On the effective manipulation of digital objects: A prototype-based instantiation approach. In *Proceedings of the 9th European Conference on Digital Libraries (ECDL 2005)*, pages 26–37, 2005.
2. T. Staples, R. Wayland, and S. Payette. The fedora project: An open-source digital object repository management system. *D-Lib Magazine*, 9(4), April 2003.

Knowledge Generation from Digital Libraries and Persistent Archives

Paul Watry[1], Ray R. Larson[2] and Robert Sanderson[1]

[1] University of Liverpool, Liverpool, L69 3DA, United Kingdom
[2] University of California, Berkeley, School of Information, 102 South Hall Berkeley, CA 94720-4600

Abstract. This poster describes the ongoing research of the Cheshire project with a particular focus on knowledge generation and digital preservation. The infrastructure described makes use of tools from computational linguistics, distributed parallel processing and storage, information retrieval and digital preservation environments to produce new knowledge from very large scale datasets present in the data grid.

1 Introduction

The University of Liverpool and the San Diego Supercomputer Center (SDSC) are jointly working on technologies and infrastructures which will support digital library services and persistent archives based on the Storage Resource Broker[1] (SRB) data grid technology. The objective of our work is to develop and implement an architecture, based on the state of the art in the data grid, persistent archive, and digital library communities, which will support all the processes within the information life-cycle: ingest, storage, management, discovery, presentation and reuse.

The novel technologies in support of this objective include the Multivalent[2] digital object management system for the preservation of digital entities and the Cheshire3 digital library architecture, which forms the infrastrucure of the UK National Centre for Text Mining (NaCTeM) amongst other services. These technologies have been integrated with the Storage Resource Broker (SRB) to provide the storage repository abstractions to enable preservation environments. This infrastructure, supporting the management of persistent archives, has the additional potential of automatically generating knowledge from the content of these large archives.

We are now in a position to investigate the application of recent advances in computational linguistics in ways which will address many of the challenges facing the information community today.

2 Background and Rationale

The challenges of generating knowledge are both compounded and facilitated by the large amount of data currently being generated by the scientific community.

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 504–507, 2006.

The requirements of dealing with this type of data reflect a range of issues relating to its long-term storage. In this respect, our knowledge generation approach will build on our current digital preservation activities and complements this with the application of new technologies deriving from the information technology and computational linguistic communities. We argue that such an aggressive integration of technologies is the best opportunity we have to address the complexity and scale of information overload which is currently facing us.

The integration of data grid, digital library, and persistent archive technologies cited above is currently ongoing for a number of projects conducted in partnership with the SDSC. These include:

- Various sub-projects of the National Science Digital Library (NSDL), including the Persistent Archive Testbed and Digital Preservation Management projects which are exploring automation of archival processes for document and multimedia collections. Also the Educational Material Categorization project, which uses multiple techniques implemented within the Cheshire3 architecture to analyze the grade level of material in the archive.
- The National Archives and Records Administration (NARA) Persistent Archives and the National Partnership for the Advancement of Computational Infrastructure (NPACI) collaboration project, the core development of which will demonstrate the automation of archival processes; develop constraint-based collection management systems; and develop the concepts of digital ontologies as a combined migration/emulation approach to digital preservation.
- The Arts and Humanities (AHDS) prototype, which has been funded as part of the Joint Information Systems Committee (JISC) Virtual Research Environment (VRE) initiative, which develops further the methodology of the above projects and, on completion, will enable secured and distributed ingestion of digital objects into a persistent archive, adding content management capabilities. For this project, the Kepler[3] workflow acts as the central point for the initial organization of data and for tracking bitstreams and metadata functionality.

Each of the above projects seeks to examine and utilize the ability of data grid technologies to meet the digital preservation requirements; to demonstrate the generality of the data grid approach; and to apply the benefits achieved by using data grid technologies in other settings.

Our approach covers the following components:

- The role of computational linguistics and ontological processes to label and characterize knowledge.
- The use of digital library technologies to apply the abstracted rules to other disciplines.
- The use of data grid technologies to provide seamless and persistent access to very large scale storage.
- The use of presentation technologies supported by the data grid to interpret digital entities.

3 Components

3.1 Computational Linguistics

The primary challenge in the characterization of knowledge is that the meaning of semantic terms within any given domain often depends on associated context: projects such as the NSDL Education Digital Library have demonstrated the possibilities of clustering semantic terms and relationships to produce knowledge and we are using these techniques with the Cheshire system to support domain analysis.

3.2 Digital Library Technologies

We will deploy digital library technologies, as implemented in the Cheshire3 system to add additional domain analysis capabilities. Cheshire3, as presented at JCDL 2005[4] and the computational and data grid components at INFOSCALE 2006[5], is an information framework based on international standards. It implements an extensible workflow system that enables easy integration of many different components and processes without significant development overheads. The architecture integrates the Multivalent document parsing capabilities along with natural language parsing tools and machine learning systems such as clustering and classification. The combined framework is therefore able to parse and categorize files that have been registered within an SRB collection.

Within the context of existing work for SDSC, the Cheshire3 system is being used to index the NSDL Educational Material Categorization Project. This uses the capability of the system to cluster together topics which may be semantically related by automating the process of association between natural language and ontologies. This capability appears to be effective in enabling users to map their query to controlled vocabularies used in descriptive metadata and may be used to cross-search different thesauri and automate associations between them and the user's query. We are further able to apply the clustering techniques to analyze content or domains through the training of document classifiers based on manual or semi-automatic classifications.

The additional inclusion of techniques discussed above, will allow us to generate and compare selected clusters of semantically labeled features. This capability forms the extraction and organization of knowledge, which will be used in the generation and application of rules to discover the existence of new relationships.

3.3 Presentation Technologies

The Multivalent digital object management system, as described at ECDL 2005, is a functioning tool which will satisfy the core requirement of preserving the ability to manipulate the encoding format of a digital entity. Using Multivalent, we are able to interpret digital entities directly from the bitstream without requiring specific supporting software or hardware. This provides us with the critical function of being able to support the original operations for manipulating digital entities.

3.4 Data Grid Technologies

The use of the above processes to generate knowledge relies heavily on the use of data management technologies developed at SDSC for persistent archives projects. While there are multiple instances of deployment for ambitious persistent archives projects (notably the NARA prototype), the evolution of the SRB data grid middleware has highlighted the need to address a number of related research activities relating to the automation of archival processes and the development of constraint-based collection management systems.

4 Conclusion

Our strategy to generate knowledge across multiple domains, as presented in this paper, combines data grid abstractions for storage (using the SRB) with presentation applications (Multivalent) and digital library and content management functionalities (Cheshire). Once we analyze the requirements for generating knowledge across diverse scientific domains, we realize that we need to guarantee the manipulation of digital entities in the future across the different infrastructures which may be used in different scientific disciplines. The work being done to facilitate digital preservation has therefore inspired a new approach to knowledge generation which will in the future make intelligent use of the massive data stores characteristic of the scientific world.

References

1. Rajasekar, A. et al. Storage Resource Broker - Managing Distributed Data in a Grid, Computer Society of India Journal, vol. 33, no. 4, pp. 4254, 2003.
2. Phelps, T., Watry, P.: A No-Compromises Architecture for Digital Document Preservation. In Procs. Research and Advanced Technology for Digital Libraries, 9th European Conference (2005), pp. 266-277
3. Ludscher, B. et al. Scientic Workow Management and the Kepler System, in Concurrency and Computation: Practice and Experience, Special Issue on Scientific Workflows, 2005
4. Larson, R., Sanderson, R.: Grid-Based Digital Libraries: Cheshire3 and Distributed Retrieval. In Procs. Joint Conferece on Digital Libaries (2005), pp. 112–114
5. Sanderson, R., Larson R.: Indexing and Searching Tera-Scale Grid-Based Digital Libraries. In Procs. First International Conference on Scalable Information Systems (2006)

Managing the Quality of Person Names in DBLP

Patrick Reuther[1], Bernd Walter[1], Michael Ley[1],
Alexander Weber[1], and Stefan Klink[2]

[1] Department of Databases and Information Systems (DBIS),
University of Trier, Germany
{reuther, walter, ley, aweber}@uni-trier.de
http://dbis.uni-trier.de

[2] Institute of Applied Informatics and Formal Description Methods,
Universitt Karlsruhe (TH), Germany
Stefan.Klink@aifb.uni-karlsruhe.de
http://www.aifb.uni-karlsruhe.de

Abstract. Quality management is, not only for digital libraries, an important task in which many dimensions and different aspects have to be considered. The following paper gives a short overview on DBLP in which the data acquisition and maintenance process underlying DBLP is discussed from a quality point of view. The paper finishes with a new approach to identify erroneous person names.

1 Introduction

The amount of information is growing exponentially. This counts also for scientific domains where one can observe a fast growth in publications. Scientific publications are the appropriate means to communicate results and new insights. Besides on a more personal level and enhanced by the often cited publish or perish mentality publications are a sort of collecting credit points for the CV. Using bibliographic statistics is more and more the first choice to evaluate scientists on an institutional level. It is obvious that all the mentioned aspects build on reliable collection, organization and access to publications.

Of utmost importance for any provider of bibliographical content is the quality of the service they offer. Quality management is ubiquitous and plays a central role in nearly any domain. For services offering access to scientific publications data quality management, a part of quality management in general, is the central challenge. Data quality comprises many different dimensions and aspects. Redman, for example, presents a variety of dimensions such as the completeness, accuracy, correctness, currency and consistency of data as well as two basic aspects to improve quality: data-driven and process driven strategies [4].

The remainder of this paper gives an overview on DBLP and its data acquisition and maintenance process, focussing on quality problems, especially problems connected to personal names. The paper ends with the presentation of a social network based approach to identify erroneous person names.

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 508–511, 2006.

2 DBLP and Quality Management

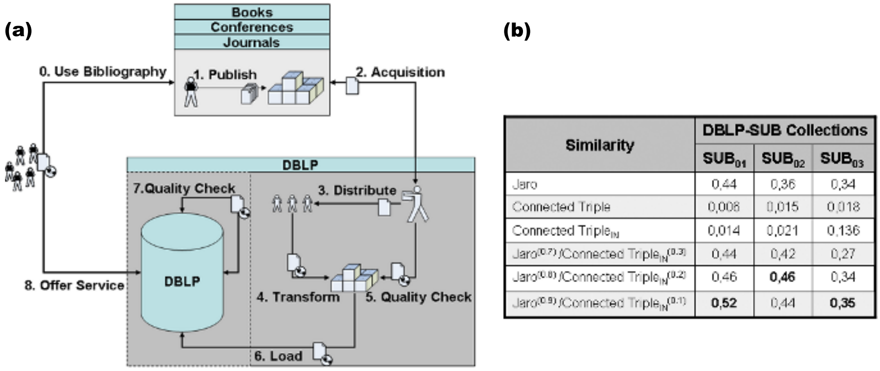
DBLP (*Digital Bibliography & Library Project*) [2] is an *internet newcomer* offering access to scientific publications. Today (May 2006) DBLP indexes more than 750.000 publications published by more than 450.000 authors.

Building a bibliographic database always requires decisions between quality and quantity. For DBLP we decided to prefer the quality of the records we offer to the quantity. It is easy to produce a huge number of bibliographic records disregarding quality aspects like standardization of journal names or person names. However, as soon as you try to guarantee that an entity is always represented by exactly the same character string and no entities share the same representation, data maintenance becomes very expensive.

Traditionally this process is called *authority control.* In DBLP the number of different journals and conference series is a few thousands so that guaranteeing consistency is not a serious problem. In contrast, authority control for person names is much harder due to the magnitude of $> 450k$ and the fact that available information is often incomplete and contradictory.

On a high level representation the data acquisition and maintenance process of DBLP shown in Fig. 1(a) can be seen as a ETL-process often found in data warehousing in which data is **E**xtracted from outside sources, **T**ransformed to fit business needs and finally **L**oaded into the database for further usage. The data of interest for DBLP which is extracted are publications authored by scientist and published in either journals, conference proceedings or more general, scientific venues. For DBLP there is a broad range of primary information sources. Usually we get electronic documents but sometimes all information has to be typed in manually. In some cases we have only the front pages (title pages, table of contents) of a journal or proceedings volume. The table of contents often contains information inferior to the head of the article itself: Sometimes the given names of the authors are abbreviated. The affiliation information for authors often is missing. Many tables of contents contain errors, especially if they were produced under time pressure like many proceedings. Even in the head of the article itself you may find typographic errors. A very simple but important policy is to enter all articles of a proceedings volume or journal issue in one step. In DBLP we make only very few exception from this *all or nothing policy.* For data quality this has several advantages over entering CVs of scientists or reference lists of papers: It is easier to guarantee complete coverage of a journal or conference series. There is less danger to become biased in favor of some person(s).

After the acquisition of data it is distributed in order to transform the data efficiently into the internal representation. Up to now, this is mainly the work of student assistants extra hired to accomplish the time consuming task of transformation. After the transformation in which first consistency checks are naturally applied the data is subject to a more thorough quality analysis. In this stage problematic cases not handled in transformation as well as further error detection are the main tasks. This work is mainly done by M. Ley for which he makes use of small tools such as the DBL-Browser [1] and scripts to automatically identify erroneous data. Here data edits are integrated into the process or



Similarity	DBLP-SUB Collections		
	SUB_{01}	SUB_{02}	SUB_{03}
Jaro	0,44	0,36	0,34
Connected Triple	0,006	0,015	0,018
Connected $Triple_{IN}$	0,014	0,021	0,136
$Jaro^{(0.7)}$ /Connected $Triple_{IN}{}^{(0.3)}$	0,44	0,42	0,27
$Jaro^{(0.8)}$ /Connected $Triple_{IN}{}^{(0.2)}$	0,46	**0,46**	0,34
$Jaro^{(0.9)}$ /Connected $Triple_{IN}{}^{(0.1)}$	**0,52**	0,44	**0,35**

Fig. 1. (a)Aquisition/Maintenance process (b)Evaluation of Avg. Precision

information chain making them a part of the process driven quality management. Example data edits in use are simple rules like firing a warning if two person names have a string distance which is smaller then a predefined threshold or if formatting conventions are not met. After the quality assurance the new data is loaded into the main DBLP database. At a typical working day we add about 500 bibliographic records. It is unrealistic to belief that this is possible without introducing new errors and without overlooking old ones. It is unavoidable that care during the input process varies. Therefore even after integrating the new records into the live system data quality is checked regularly. This data driven quality management is again supported by simple scripts and small tools. The loop of the data acquisition and maintenance process for DBLP closes when researchers use the system, especially the new entered bibliographical records and use them to produce new publications which some day will most likely be integrated into the DBLP system. From a data quality point of view improvements for data quality can only be made for the stages (2. Acquisition) to (7. Quality Check). The primary information creation and publishing (1. Publishing) is not in our area of responsibility and therefore can not be ameliorated, although improvements such as implementing an International Standard Author Number (ISAN), analogously to the ISBN known for publications, would confine the problems connected to names dramatically [5].

3 Personal Name Matching with Co-author Networks

From reflecting on how we find errors and inconsistencies concerning person names, we designed new similarity measurements based on a co-author network G in which authors are represented as vertices V, and co-authorship builds the edges E. For two person names a simple way to determine their similarity is to count the amount of Connected Triples they are part in. A Connected Triple $\wedge = \{V_\wedge, E_\wedge\}$ can be described as a subgraph of G consisting of three vertices with $V_\wedge = \{A_1, A_2, A_3\} \subset V$ and $E_\wedge = \{e_{A_1,A_2}, e_{A_1,A_3}\} \in E, \{e_{A_2,A_3}\} \notin E$. The

Connected Triple similarity of two names i and j is then calculated by $\frac{|C_{\wedge_{ij}}|}{|C_{\wedge_{max}}|}$ where $|C_{\wedge_{ij}}|$ is the number of Connected Triples between i and j and $|C_{\wedge_{max}}|$ the maximal number of Connected Triples between any two authors. This simple similarity function can be systematically improved by considering the amount of publications which lead to the number of Connected Triples as well as the distribution of authors in these publications. Therefore the edges in the co-author network will be weighted according to Liu et al. [3]. With $V = \{v_1, \ldots, v_n\}$ as the set of n authors, m the amount of publications $A = \{a_1, \ldots, a_k, \ldots a_m\}$ and $f(a_k)$ the amount of authors of publications a_k the weight between two authors v_i and v_j for publications a_k is calculated by $g(i, j, k) = \frac{1}{f(a_k)-1}$. Thereby the weight between two authors for one publication is smaller the more authors collaborated on this publication. Considering the amount of publications two authors i and j collaborated on together, an edge between these authors is calculated with $c_{ij} = \sum_{k=1}^{m} g(i, j, k)$ which leads to higher weights the more publications the two authors share. Applying a normalisation the weight between two authors i and j considering the amount of co-authors and publications is calculated by $w_{ij} = \frac{c_{ij}}{\sum_{r=1}^{n} c_{ir}}$ leading to a directed co-author graph. The similarity of two authors using Connected Triples can consequently be either calculated on incoming edges ($ConnectedTriple_{in} = \sum_{\forall c \in V with\ e_{ci}, e_{cj} \in E, e_{ij} \notin E} w_{ci} + w_{cj}$) or outgoing edges ($ConnectedTriple_{out} = \sum_{\forall c \in V with\ e_{ic}, e_{jc} \in E, e_{ij} \notin E} w_{ic} + w_{jc}$). Evaluations (see Fig. 1(b)), show that the sketched approaches lead to a reasonable precision especially when combined with syntactical criteria.

4 Conclusion

Managing data quality plays an increasing role for service providers in the internet such as DBLP which offers access to bibliographic records. The most time consuming quality task is the task of offering consistent data. Especially person names are error prone and hard to deal with. To confine the problems connected with person names we constantly develop new similarity measures which we evaluate on a specially designed framework making use of new test collections. The promising approaches are then integrated into the data acquisition and maintenance process of DBLP to guarantee a high quality of the data.

References

1. S. Klink, M. Ley, E. Rabbidge, P. Reuther, B. Walter, and A. Weber. Browsing and visualizing digital bibliographic data. In *VisSym 2004*, pages 237–242, 2004.
2. M. Ley. The dblp computer science bibliography: Evolution, research issues, perspectives. In *SPIRE*, pages 1–10, 2002.
3. X. Liu, J. Bollen, M. L. Nelson, and H. V. de Sompel. All in the family?, 2005. online: http://public.lanl.gov/liu_x/trend.pdf.
4. T. C. Redman. *Data Quality for the Information Age.* Artech House, 1996.
5. P. Reuther. Personal name matching: New test collections and a social network based approach. *Universität Trier, Technical Report*, 06-01, 2006.

MedSearch: A Retrieval System for Medical Information Based on Semantic Similarity

Angelos Hliaoutakis[1], Giannis Varelas[1], and Euripides G.M. Petrakis[1], and Evangelos Milios[2]

[1] Dept. of Electronic and Computer Engineering
Technical University of Crete (TUC)
Chania, Crete, GR-73100, Greece
angelos@softnet.tuc.gr, varelas@softnet.tuc.gr,
petrakis@intelligence.tuc.gr

[2] Faculty of Computer Science, Dalhousie University
Halifax, Nova Scotia, B3H 1W5, Canada
eem@cs.dal.ca

Abstract. *MedSearch*[1] is a complete retrieval system for Medline, the premier bibliographic database of the U.S. National Library of Medicine (NLM). *MedSearch* implements *SSRM*, a novel information retrieval method for discovering similarities between documents containing semantically similar but not necessarily lexically similar terms.

1 Introduction

MedSearch is a complete retrieval system for medical literature. It supports retrieval by *SSRM* (*Semantic Similarity Retrieval Model*) [1], a novel information retrieval method which is capable for associating documents containing semantically similar (but not necessarily lexically similar) terms. *SSRM* suggests discovering semantically similar terms in documents and queries using term taxonomies (ontologies) and by associating such terms using semantic similarity methods (e.g., [2]). *SSRM* demonstrated very promising performance achieving significantly better precision and recall than Vector Space Model (VSM) for retrievals on Medline.

2 Semantic Similarity Retrieval Model (*SSRM*)

As it is typical in information retrieval, documents are represented by term vectors and each term is initially represented by its $tf \cdot idf$ weight. For short queries specifying only a few terms the weights are initialized to 1. Then, *SSRM* works in three steps:

Term Re-Weighting: The weight q_i of each query term i is adjusted based on its relationships with semantically similar terms j within the same vector

[1] http://www.intelligence.tuc.gr/medsearch

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 512–515, 2006.

$$q_i = q_i + \sum_{\mathrm{sim}(i,j) \geq t}^{j \neq i} q_j \mathrm{sim}(i,j), \tag{1}$$

where t is a user defined threshold ($t = 0.8$ in this work). Semantic similarity between terms is computed according to the method described in [2]. Multiple related terms in the same query reinforce each other (e.g., "train", "metro"). The weights of non-similar terms remain unchanged (e.g., "train", "house").

Term Expansion: First, the query is augmented by synonym terms, using the most common sense of each query term. Then, the query is augmented by semantically similar terms higher or lower in the taxonomy (i.e., hypernyms and hyponyms). The neighborhood of the term in the taxonomy is examined and all terms with similarity greater than threshold T ($T = 0.9$ in this work) are also included in the query vector. This expansion may include terms more than one level higher or lower than the original term. Then, each query term i is assigned a weight as follows

$$q_i' = q_i + \sum_{\mathrm{sim}(i,j) \geq T \text{ and } j \in Q}^{i \neq j} \frac{1}{n} q_j \mathrm{sim}(i,j), \tag{2}$$

where n is the number of hyponyms of each expanded term j, q_i is the weight of term i before expansion and Q is the subset of the set of original query terms that led into new terms added to the expanded query. For hypernyms $n = 1$. Notice that $q_i = 0$ if term i was not in the original query but was introduced during the query expansion process. It is possible for a term to introduce terms that already existed in the query. It is also possible that the same term is introduced by more than one other terms. Eq. 2 suggests taking the weights of the original query terms into account and that the contribution of each term in assigning weights to query terms is normalized by the number n of its hyponyms. After expansion and re-weighting, the query vector is normalized by document length, like each document vector.

Document Similarity: The similarity between an expanded and re-weighted query q and a document d is computed as

$$Sim(q,d) = \frac{\sum_i \sum_j q_i d_j \mathrm{sim}(i,j)}{\sum_i \sum_j q_i d_j}, \tag{3}$$

where i and j are terms in the query and the document respectively. The similarity measure above is normalized in the range [0,1]. Notice that, if there are no semantically similar terms ($\mathrm{sim}(i,j) = 0 \ \forall i \neq j$) *SSRM* is reduced to VSM.

Expanding and re-weighting is fast for queries, which are typically short, consisting of only a few terms, but not for documents with many terms. The method suggests expansion of the query only. Notice that expansion with low threshold values T (e.g., $T = 0.5$) is likely to introduce many new terms and diffuse the topic of the query (topic drift).

3 *MedSearch*

Medline[2] is the premier bibliographic database of the U.S. National Library of Medicine (NLM), indexing more that 15 million references (version 2006) to journal articles in life sciences, medicine and bio-medicine. In addition to title, abstract and authors, Medline stores a rich set of metadata associated with each article such as language of publication, publication type, dates, source of publication and relations between articles. Articles in Medline are also indexed (manually by experts) by a set of descriptive MeSH terms.

MedSearch supports retrieval of bibliographic information on Medline by VSM as well as by semantic retrieval by *SSRM* using MeSH[3] as the underlying reference ontology. VSM and *SSRM* are implemented on top of Lucene[4] a full-featured text search engine library in Java. All documents are indexed by title, abstract and MeSH terms. These descriptions are syntactically analyzed and reduced into separate vectors of MeSH terms which are matched against the queries according to Eq. 3 (as similarity between expanded and re-weighted vectors). The weights of all MeSH terms are initialized to 1 while the weights of titles and abstracts are initialized by $tf \cdot idf$. The similarity between a query and a document is computed as

$$Sim(q,d) = Sim(q, d_{MeSH-terms}) + Sim(q, d_{title}) + Sim(q, d_{abstract}), \quad (4)$$

where $d_{MeSH-terms}$, d_{title} and $d_{abstract}$ are the representations of the document MeSH terms, title and abstract respectively. This formula suggests that a document is similar to a query if most of its components are similar to the query.

The specification of threshold T in *SSRM* may depend on query scope or user uncertainty. A low value of T is desirable for broad scope queries or for initially resolving uncertainty as to what the user is really looking for. The query is then repeated with higher threshold. A high value of T is desirable for very specific queries: Users with high degree of certainty might prefer to expand with a high threshold or not to expand at all. The option is also adjustable at the interface of *MedSearch*. In this work we set $T = 0.9$ (i.e. the query is expanded only with very similar terms).

A set of 15 of medical queries were prepared by an independent medical expert. Each query specified between 3 and 10 terms and retrieved the best 20 answers. The results were evaluated by the same medical expert. Each method is represented by a *precision/recall* curve. Each point on a curve is the average precision and recall over all queries. Fig. 1 indicates that VSM with query expansion is obviously the worst method. Each query term is augmented by its "Entry Terms" in MeSH (i.e., general related terms which are not always synonyms). Notice that no exact synonymy relation is defined in MeSH. For this reason, in *SSRM* we do not apply expansion with Entry Terms. However, query terms are expanded with semantically similar terms in the neighborhood of each term

[2] http://www.ncbi.nlm.nih.gov/entrez

[3] http://www.nlm.nih.gov/mesh

[4] http://lucene.apache.org

according to Eq. 2. Semantic information retrieval by *SSRM* is more effective than classic information retrieval by VSM achieving up to 20% better precision and up to 20% better recall.

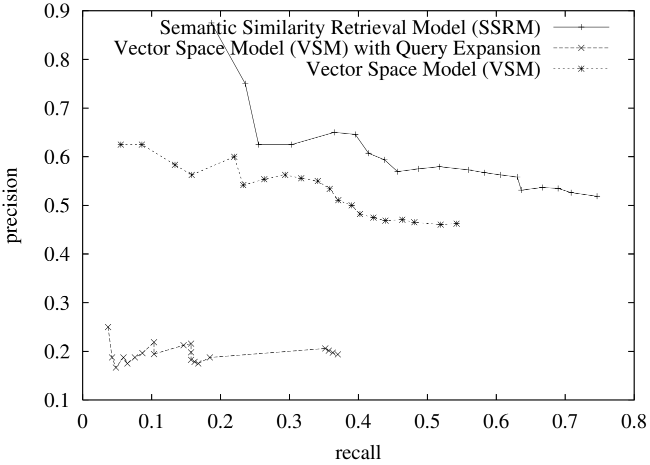


Fig. 1. Precision-recall diagram of *SSRM* and VSM for retrievals on Medline

4 Conclusions

MedSearch is an information retrieval system for medical literature and is accessible on the Web. *MedSearch* supports retrieval by VSM (the classic retrieval method) and by *SSRM*. *SSRM* demonstrates promising performance improvements over VSM. *SSRM* can work in conjunction with any taxonomic ontology (e.g. WordNet).

Acknowledgement

We thank Dr. Qi, Dalhousie University, for her help in preparing the set of medical queries. This work was supported by project BIOPATTERN (FP6-NoE, Project No. 508803) of the European Union (EU).

References

1. Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E., Milios, E.: Semantic Similarity Methods in WordNet and their Application to Information Retrieval on the Web. In: 7^{th} ACM Intern. Workshop on Web Information and Data Management (WIDM 2005), Bremen, Germany (2005) 10–16
2. Leacock, C., Chodorow, M.: Combining Local Context and WordNet Similarity for Word Sense Identification in WordNet. In Fellbaum, C., ed.: An Electronic Lexical Database. MIT Press (1998) 265–283

Metadata Spaces: The Concept and a Case with REPOX

Nuno Freire and José Borbinha

INESC-ID, Rua Alves Redol 9, Apartado 13069,
1000-029 Lisboa, Portugal
nuno.freire@bn.pt, jlb@ist.utl.pt

Abstract. This paper describes REPOX, an XML infrastructure to store and manage metadata, in the sense it is commonly defined in digital libraries. The purpose is to make it possible, in alignment with an Enterprise Architecture model, to develop a component of a Service Oriented Architecture that can manage, transparently, large amounts of descriptive metadata, independently of their schemas or formats, and for the good of other services. The main functions of this infrastructure are submission (including synchronisation with external data sources), storage (including long-term preservation) and retrieval (with persistent linking). The case is demonstrated with a deployment at the National Library of Portugal, using metadata from two information systems and three schemas: bibliographic and authority data from a union catalogue and descriptive data from an archival management system.

1 Introduction

This paper uses the concept of Metadata Space to propose the design and deployment off a corresponding solution for the National Library of Portugal (BN), named REPOX. REPOX is aligned with the concept of dataspace, as defined in [3]. It is also an important element to support the actual analysis process in place at BN toward the definition of a real Enterprise Architecture (EA[1]) and the development of a computing environment according to the principles of a Service Oriented Architecture (SOA)[2]. The purpose of a Metadata Space is, in a SOA, to manage, transparently, large amounts of descriptive metadata, independently of their schema or format, and for the good of other services. The main functions of this infrastructure are submission (including synchronisation with external data sources), storage (including long-term preservation) and retrieval (with persistent linking).

This paper follows with a description of the design and deployment of the Metadata Space at BN, followed by some conclusions and future work.

2 The REPOX Data Model

REPOX is intended to manage, as a first objective, the storage in XML of records originated from multiple data sources of one organization. It maintains a temporal

[1] http://en.wikipedia.org/wiki/Enterprise_Architect

[2] http://en.wikipedia.org/wiki/Service-oriented_architecture

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 516 – 519, 2006.

history of all the versions submitted of every record, with its respective date, all coded in XML.

The way the records are coded in REPOX addresses the preservation requirements in OAIS [4]. It is based on non proprietary tools for the digital preservation, is robust and flexible, provides mechanisms for self-description and validation of digital resources, and is not application or platform based [1] [2]. The data model actually defined for REPOX is shown in Fig. 1.

Internally, REPOX organizes the metadata sets as data collections which have an associated interface with the data source. These interfaces are software components responsible for obtaining the records, coding them according their XML schema, and submitting them to REPOX.

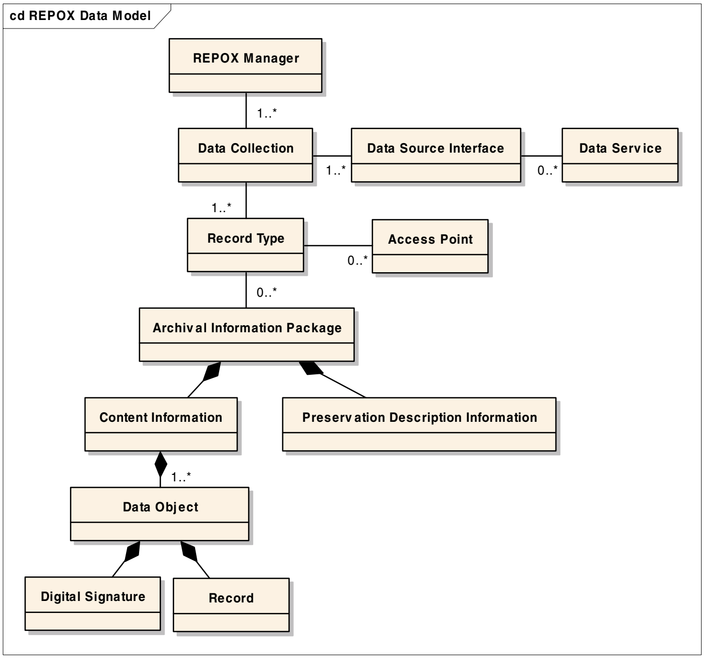


Fig. 1. The REPOX data model

REPOX associates with each data collection a set of record types. These represent the entities in the data collection that will be collected from the data source in a XML record representation and managed by REPOX. In order to enable the records to be retrieved by more than their identifier, REPOX requires information about the access points. Therefore, the definition of a record type has associated access points. These are used to create indexes, for efficient searching or access. Access points may also organize records in collections, a feature that is particularly useful in the scope of libraries.

A data source interface can be associated to a set of services. These are a special kind of services that run inside REPOX, having direct access to the repository.

Each record managed by REPOX is wrapped in an AIP - Archive Information Package, as defined by the OAIS model. It is coded according to the REPOX AIP schema and stored in a file on the file system. Since file systems security against changes in the files is not enough to guarantee the level of authenticity needed on the long term, a digital signature is kept with each version of the record.

REPOX maintains a URN space for the identification of data collections, record types, AIPs and each version of the records.

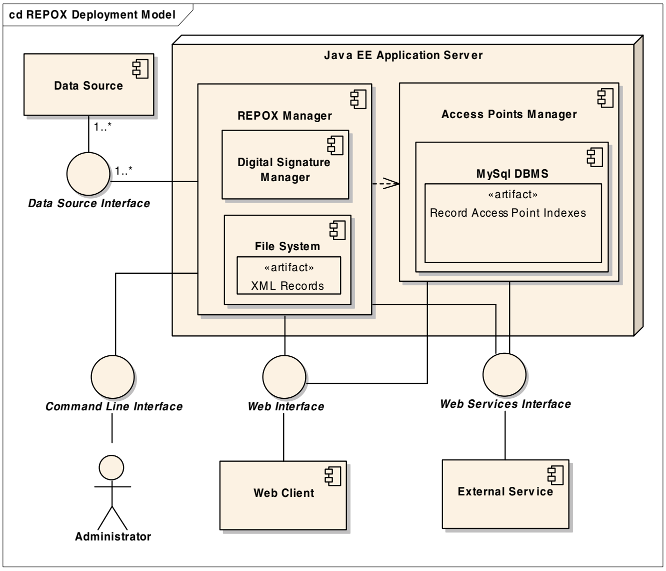


Fig. 2. A REPOX deployment

3 The REPOX Architecture

The main component of the REPOX infrastructure is the REPOX Manager. It is a Java EE application that implements de model described in the previous section. Fig. 2 represents a deployment of REPOX with an arbitrary number of data sources.

The XML repository, via the data source interfaces, receives new records periodically from the data sources. These records are wrapped in a XML AIP and archived in the file system along with a digital signature. For an efficient retrieval of the records, the REPOX manager has an Access Point Manager.

A web interface and several command line tools are available for the management of the repository, as also a web services interface for external systems

4 REPOX in Use at BN

A REPOX infrastructure is in production in the BN. Currently, it is supporting two data collections: PORBASE - the national union catalogue, and the Archive of the Contemporary Portuguese Culture (ACPC, a department of the BN). This REPOX dataspace of PORBASE contains nearly 3 million records, comprising nearly 26

GBytes. It maintains 33 access points, and manages 21 collections of bibliographic records and 4 of authority records.

The data collection of the ACPC contains archival records in a format that follows the ISAD(G) [3] rules, and also authority records according to the ISAAR(CPF)[4].

Several services at BN are already in stable production using REPOX, such as:

- **OAI-PMH[5] Service** (http://oai.bn.pt): This service makes available records from PORBASE to external services. A regular client of this service is, for example, the portal TEL - The European Library (http://www.theeuropeanlibrary.org/)
- **Google Scholar**: PORBASE is one of the bibliographic databases pioneer in providing its data to Google Scholar for indexing. Portuguese users of Google Scholar can see now links to the OPAC of PORBASE when their search hits one of its records. The mechanism for the sharing of records is built over REPOX.

5 Conclusions and Future Work

REPOX started its operation with PORBASE on the 1st of October of 2005. Immediately afterwards, it was noticeable the increase in the availability and performance of access to the data, carried out by several departments of the National Library and the partners' libraries. As of now, that represents a total 3 million records, in their way preserved independently of any specific software or hardware.

The usage of the REPOX system on real scenarios has demonstrated its usefulness and efficiency, but also has demonstrated weaknesses. Future developments will focus on the linking of records and improvements to the retrieval mechanisms. Also, the work on the Long-term Archive Protocol [5] is being followed and its implementation is being considered.

References

1. Electronic Resource Preservation and Access Network (ERPANET): Urbino Workshop: XML for Digital Preservation (2002) http://eprints.erpanet.org/archive/00000002/01/UrbinoWorkshopReport.pdf
2. Boudrez, F.: XML and electronic record-keeping (2002) http://www.expertisecentrumdavid.be/davidproject/teksten/XML_erecordkeeping.pdf
3. Franklin, M., Halevy, A., Maier, D.: From Databases to Dataspaces: A New Abstraction for Information Management. ACM SIGMOD Record (2005)
4. Consultative Committee for Space Data Systems. OAIS - Reference Model for an Open Archival Information System (2002)
5. Blazic, A.J., Sylvester, P., Wallace, C.: Long-term Archive Protocol (LTAP) (2006) http://www.ietf.org/internet-drafts/draft-ietf-ltans-ltap-01.txt

[3] http://www.ica.org/biblio/cds/isad_g_2e.pdf

[4] http://www.ica.org/biblio/isaar_eng.html

[5] http://www.openarchives.org/

Multi-Layered Browsing and Visualisation for Digital Libraries

Alexander Weber[1], Patrick Reuther[1], Bernd Walter[1],
Michael Ley[1], and Stefan Klink[2]

[1] Department of Databases and Information Systems (DBIS),
University of Trier, Germany
{aweber, reuther, walter, ley}@uni-trier.de
http://dbis.uni-trier.de

[2] Institute of Applied Informatics and Formal Description Methods,
Universität Karlsruhe (TH), Germany
Stefan.Klink@aifb.uni-karlsruhe.de
http://www.aifb.uni-karlsruhe.de

Abstract. For a scientific researcher it is more and more vital to find relevant publications with their correct bibliographical data, not only for accurate citations but particularly for getting further information about their current research topic.

This paper describes a new approach to develop user-friendly interfaces: *Multi-Layered-Browsing.* Two example applications are introduced that play a central role in searching, browsing and visualising bibliographical data.

1 Introduction

The widespread use of computers for acquisition, production, and archiving of documents lead to more and more information in electronic form. The ease with which documents are produced and shared has lead to an exponential increase of information reachable by each user. More than 40 years ago, Maron and Kuhns predicted that indexed scientific information will double every 12 years [4]. Even in the academical area, new conferences, journals, and other publications are appearing quickly and they increase the huge amount of existing scientific publications in an alarming manner, e. g., Cleverdon estimates the amount of publications of the most important scientific journals to 400,000 per year [1] and INSPEC, the leading English-language bibliographic information service, is growing at the rate of 350,000 records each year.

Finding relevant information within these masses of data is a challenging task. Particularly, in query based information systems like Yahoo, Lycos or Google users have severe problems even to formulate exact queries.

For the digital library domain where bibliographical data is the central information we developed a more user-friendly and efficient way for searching and browsing through bibliographical data by a combination of a query-based and browsing-based approach. Starting from an unspecific query users can browse

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 520–523, 2006.

through the bibliographical data by clicking items like authors, title words, conferences etc. As already indicated by [2] that users like the textual as well as the graphical nature of information representation we enhanced the browsing-based approach by the so called *Multi-Layered-Browsing*. During the browsing process all data is visualised in a textual way and in parallel by appropriate graphical techniques which enables users to better understand their search domain and consequently offers the opportunity to get an overview of their information need very quickly. Furthermore, it helps them find relevant authors or publications and above all provides information about further researchers, important conferences or journals. The following chapters introduce the *DBL-Browser* and the *ML-Browser* which support a searching and browsing-based approach within the bibliographical data.

2 Multi-Layered-Browsing

Most modern library systems offer only a very limited freedom to the user. There are only some very strict ways to search and to browse the data set. Most of the time a simple search mask is offered and the user has to browse through the results in a mainly linear way. We are offering a better approach where the user has multiple possibilities to go different ways on his search path. He does not need to follow a strict path, but is flexible to choose the way that leads him most efficiently to his goal. We call this concept *Multi-Layered-Browsing*: The user can – metaphorically speaking – change between different layers during his way on his search path. These layers are always offering different possibilities to navigate from the current point to the next. To help the user even more, the layers do not only provide textual representations of the current visible data subset, but use different graphical ways, so the user can choose what representation is best for his current information need. An example can be seen in fig. 1 with three different 'layers' for a given author. The first layer (top-right) shows a textual representation of the author's publications, the second (bottom-left) shows the data as a histogram and the third layer (bottom-right) shows the co-author relation as a graph. Our approach to support the user is to develop user-friendly applications, that focus on the user's information need. The following sections introduce two such tools: The *Multi-Layer-Browser* and the *DBL-Browser*.

2.1 The Multi-Layer-Browser

For testing this Multi-Layered-Browsing we developed the *ML-Browser* – a tool that always shows different layers of the current data subset (see fig. 1). It is designed in such a way, that it is very easy to exchange the different visualisations with own implementations or ideas. The *author-view* seen in fig. 1 can use a global *filter-object* that can be used for all layers, to only show the current focus of the user. That is, the user can for example select the year 2005 to filter out all co-authors that have not published in 2005 and to see a detailed histogram of the publications in 2005. If such a filter is not enough for shrinking the focus it

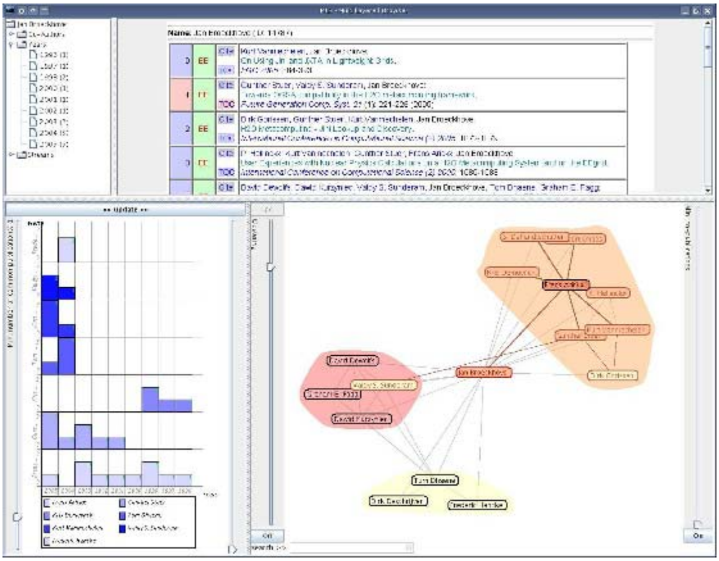

Name
EE
TOC
Future Generation Comp. Syst. 21 (1): 221-226 (2005)
Dirk Gorissen, Gunther Stuer, Kurt Vanmechelen, Jan Broeckhove
Dawid Dewolfs, Dawid Kurzyniec, Vaidy S. Sunderam, Jan Broeckhove, Tom Dhaene, Graham E. Fagg
Jan Broeckhove
Tom Dhaene
Update
On

Fig. 1. Overview of the *ML-Browser*

is possible to combine several filters. The visualisations make use of these filters as well and can so provide a specially focused view of the data.

On the one hand the combination of layered browsing and chained filtering offers the user a lot of ways to access the information he is searching for very fast; on the other hand, the user can use these techniques to get a very good overview of the surroundings of the currently viewed data, thus enabling him to browse the social networks that exist in the data and helping him to broaden his horizon.

2.2 *DBL-Browser*

The *DBL-Browser* was originally developed to browse textual visualisations of the DBLP. As the browser has evolved, so have the textual visualisations. The browser now includes, in addition to author pages, textual visualisations of search results in tabular form, BibTeX pages, and Table of Contents (TOC) pages [3]. These BibTeX and TOC pages allow users to continue to explore the knowledge domain to find other relevant documents or related authors. The author pages themselves have also evolved, allowing for a clear, consistent layout of information across all textual visualisations. Besides textual visualisations we make use of user-friendly graphical visualisations, that support users in finding the desired information. There are several things that make the *DBL-Browser* an *easy-to-use* application. One of the main aspects is it's straightforward user interface. Everybody using a common web-browser should be able to use the *DBL-Browser*. All essential features are at hand – like searching and filtering the data. The search system has all typical functions, with additional features such as combined searches or vague searches. The other main feature of the browser is the additional navigation provided by the *everything-is-clickable* concept. So there is a link on nearly every information shown in the browser – there are links to other authors, title-word searches, links to BibTeX pages, conferences, journals, or to the full-text electronic editions.

3 Conclusion and Encouragement

In this work a new approach for multi-layered browsing and visualisation for digital libraries is introduced. In contrast to query-based information systems like Yahoo, Lycos, or Google our *ML-Browser* and in particular the *DBL-Browser* are user-friendly every-day applications which help the user to find relevant publications and to browse through the bibliographical data by clicking items in various levels. On basis of a high quality database the applications offer users a multi-layered browsing environment with support of various visualisations. Textual representations, histograms and relationship graphs give a comprehensive overview of the current research area and relationships between authors. So interesting researchers can be found and contacted for exchanging information and initiating further cooperations. The Multi-Layered-Browsing approach is aswell predestined for analysing social networks regarding both author networks and conference/journal networks. This scientific area offers a lot of opportunities for further research.

Our intention is to provide the *DBL-Browser* as an open framework for experiments. Due to its modularisation, it is easy for anyone who is interested to integrate own ideas and algorithms. The XML and compressed version of the DBLP data and the source code of the browser are available on our web server: `http://dbis.uni-trier.de/DBL-Browser/` Feedback and further ideas are also very welcome. Unfortunately it is far beyond our resources to include all publications within the DBLP we are asked to consider. But we hope to find more sponsors...

References

1. C. W. Cleverdon. Optimizing convenient online access to bibliographic databases. *Information Services and Use*, 4:37–47, 1984.
2. Y. Ding, G. G. Chowdhury, S. Foo, and W. Qian. Bibliometric information retrieval system (BIRS): A web search interface utilizing bibliometric research results. *Journal of the American Society for Information Science*, 51(13):1190–1204, 2000.
3. S. Klink, M. Ley, E. Rabbidge, P. Reuther, B. Walter, and A. Weber. Browsing and visualizing digital bibliographic data. In O. Deussen, C. D. Hansen, D. A. Keim, and D. Saupe, editors, *VisSym*, pages 237–242. Eurographics Association, 2004.
4. M. E. Maron and J. L. Kuhns. On relevance, probabilistic indexing and information retrieval. *Journal of the Association for Computing Machinery (JACM)*, 7(3):216–244, July 1960.

OAI-PMH Architecture for the NASA Langley Research Center Atmospheric Science Data Center

Churngwei Chu[1], Walter E. Baskin[1], Juliet Z. Pao[1], and Michael L. Nelson[2]

[1]NASA Langley Research Center, Hampton VA, USA
[2]Old Dominion University, Norfolk VA, USA
{c.chu, w.e.baskin, j.z.pao}@larc.nasa.gov, mln@cs.odu.edu

Abstract. We present the architectural decisions involved in adding an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) interface to the NASA Langley Research Center Atmospheric Science Data Center (ASDC). We review four possible implementation strategies and discuss the implications of our choice. The ASDC differs from most OAI-PMH implementations because of its complex data model, large size (1.3 petabytes) of its Earth Science data holdings and its rate of data acquisition (>20 terabytes / month).

1 Introduction

The National Aeronautics and Space Administration (NASA) Langley Research Center Atmospheric Science Data Center (ASDC) [1] supports 42 science projects with over 1700 data sets and 2M data granules in a combination of 1.3 petabytes of online and nearline storage. ASDC is one of 8 NASA Distributed, Active Archive Centers (DAACs) in the U.S. that provide curation of federally-funded Earth Science data. The DAACs are arranged by discipline; ASDC's data sets involve radiation budget, clouds, aerosols and tropospheric chemistry. These data sets were produced to increase academic understanding of the natural and anthropogenic perturbations that influence the global climate change. In addition to archiving, distributing and processing data, ASDC also distributes metadata to other trading partners. To increase visibility of its holdings and facilitate more automated interchange with data partners, a pilot project was implemented for providing an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [2] interface for the ASDC.

2 OAI-PMH Architecture Options

The largest difficulty faced was mapping the ASDC data model into the relatively simple OAI-PMH data model of resource/item/record. Although the data model is fully discussed in [2], we present a highly summarized review here. Resources are the objects of interest and exist outside of the OAI-PMH data model; they are the "stuff" the metadata is "about". Items are in the data model and have a unique OAI identifier; they represent all the metadata records that describe a resource. Set level information is attached to the item. Items have 1 or more records. Records are

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 524 – 527, 2006.

metadata in a particular format ("metadataPrefix") and inherit their identifier from the item. Records are the unit of discourse in OAI-PMH transactions.

Conventional bibliographic applications have simple data models. For example, the resource is a book, the item uses an ISBN for its unique identifier, and the metadata is available natively in MARC as well as a Dublin Core (DC) [3] subset intended for general resource discovery. ASDC's resources are not that uniform; they have a hierarchy of "project/collection/granule" (in descending order of magnitude and rate of change) to specify provenance and versioning. For example, the International Satellite Cloud Climatology Project (ISCCP) has 9 collections, one of which (ISCCP_B3_NAT) has more than 0.5M granules. Project level metadata describes the project's purpose, instruments used and spatial/temporal coverage. Collection level metadata includes descriptions for scientific keywords and platforms such as satellites and airplanes. Granule level metadata describes only spatial/temporal coverage. In the ASDC, there are approximately 1000 record/collection metadata records and 2 million granule metadata records. Four main options were considered for mapping projects/collections (P/C) and granules into OAI-PMH:

1. Three separate repositories and corresponding baseURLs are created. The repositories are separated by how they define items: projects, collections or granules. This option maintains uniformity within each repository, but each repository must be harvested separately to acquire all the data.
2. Granules are defined as items and a single repository/baseURL is used. Separate metadataPrefixes are used to convey project and collection level metadata. This option promotes granules to the primary focus of the repository, but would result in significant duplication of project and collection metadata records because of the imbalance between granules and projects/collections: a single project metadata record is likely to be associated with thousands of granules.
3. Similar to #2, granules are defined as items and a single repository/baseURL is used. But in this case, project and collection metadata exists outside of the OAI-PMH framework. For example, since there are so few project and collection records and are likely to change very infrequently, they could be simply referenced as external URLs from the granule metadata records. This option has the advantage of uniformity (all items are granules), but not all information is directly OAI-PMH accessible.
4. The final option considered also uses a single repository/baseURL, but items are projects, collections and granules. The nature of a metadata record can be inferred from its metadataPrefix, identifier and set information. This option does not have a uniform concept of items, but the items are highly interrelated and all ASDC metadata records are accessible from a single baseURL.

After careful consideration, we decided to implement option #4. The records are differentiated by their metadataPrefix: project and collection records are available in DC and Directory Interchange Format (DIF) [4], whereas granules are only available in DIF. Although this is not compliant with the OAI-PMH since there is not a DC representation for every record, this approach was chosen because we considered the metadataPrefix as an indicator of intention. DIF is a highly specialized science data format harvested and likely to be understood only by known trading partners active in Earth Science research. We interpret a request for DC as an indication of non-expert usage, and thus harvesters requesting DC receive only the project and collection based records, which are suitable for cross-domain service providers.

The project/collection/granule hierarchy is incorporated in both identifiers and sets. This identifier specifies a granule 445025 in the collection ISCCP_B3_NAT (124th revision), and the set value describes the membership of the granule.

Identifier: oai:asdc.larc.nasa.gov:ISCCP:ISCCP_B3_NAT_124:445025
Set: info:asdc.larc.nasa.gov:ISCCP:ISCCP_B3_NAT

Similarly, the corresponding collection and project identifiers and set values would be, respectively:

Identifier: oai:asdc.larc.nasa.gov:ISCCP:ISCCP_B3_NAT
Set: info:asdc.larc.nasa.gov:ISCCP:ISCCP_B3_NAT

Identifier: oai:asdc.larc.nasa.gov:ISCCP
Set: info:asdc.larc.nasa.gov:ISCCP

Notice that for projects and collections, the set values match the identifiers. It also worth noting that in OAI-PMH, the ":" character is recursive. So requests of the form:

?verb=ListRecords&metadataPrefix=DIF&set= info:asdc.larc.nasa.gov:ISCCP

Would return all project, collection and granules records in the DIF metadata format. This request specifies DC as the metadataPrefix, so it would return metadata for projects and collections, but not metadata:

?verb=ListRecords&metadataPrefix=oai_dc&set= info:asdc.larc.nasa.gov:ISCCP

This request would return only the project metadata in DIF:

?verb=GetRecord&metadataPrefix=DIF&identifier=oai:asdc.larc.nasa.gov:ISCCP

3 Future Work and Conclusions

There are several implications for adding an OAI-PMH interface to the ASDC. First, it will result in much greater exposure of ASDC collection. We plan to expose the project and collection records to the NASA Technical Report Server (NTRS) [5]. NTRS is a "one-stop shop" for NASA authored publications, and since NTRS already uses OAI-PMH to harvest from other NASA institutional repositories, adding the ASDC collection will be easy. Furthermore, since Google supports OAI-PMH, it will be easy to expose project and collection metadata records to Google as well. Increased coverage in additional services such as NTRS and Google is an example of the "inverted repository" model in which the data objects themselves are exposed and harvested by many services and point back to their main home page. This is contrast to the more conventional model where home pages (such as [1]) are the only resource indexed and the data objects are discovered only through the services at the home page.

Our future plans also include moving from metadata harvesting to actual resource harvesting. This involves bringing the resource into the realm of the OAI-PMH data model by encoding in a complex object format (e.g., MPEG-21 Digital Item Declaration Language (DIDL) or Metadata Encoding and Transmission Standard (METS)) and treating the resulting object as a metadata format. This approach has been shown to allow for accurate repository synchronization using off the shelf harvester software [6].

ASDC is in the process of testing the repository implementation, currently based on OAICat [7], and will make the URL initially available to selected trading partners. We are working with potential partners to establish OAI-PMH use in the Earth Science community. Our implementation currently violates the letter of the OAI-PMH specification by not returning DC records for all items. Since the conversion of DIF-to-DC for granules is very lossy, we adopted this approach to prevent the likelihood of unwary harvesters acquiring millions of DC records with little value. If we choose to become technically compliant in the future, we could not expose granule identifiers for harvesters that do not provide some well-known user id and password (e.g., "earth" and "science").

We plan to use OAI-PMH to facilitate interchange between ASDC and other partners, to encourage the development of new, specialized services based on Earth Science data, and to increase the exposure of our Earth Science data holdings through increased search engine and service provider coverage.

References

1. Atmospheric Science Data Center. http://eosweb.larc.nasa.gov/
2. C. Lagoze, H. Van de Sompel, M. L. Nelson and S. Warner. *The Open Archives Initiative Protocol for Metadata Harvesting Version 2.0.* http://www.openarchives.org/OAI/openarchivesprotocol.html
3. S. Weibel, J. Kunze, C. Lagoze and M. Wolf. *Dublin Core Metadata for Resource Discovery.* Internet RFC-2413, 1998.
4. *Directory Interchange Format (DIF) Writer's Guide. Version 9.4.* 2005. http://gcmd.nasa.gov/User/difguide/
5. M. L. Nelson, J. R. Calhoun, and C. E. Mackey. "The OAI-PMH NASA Technical Report Server." *Proceedings of JCDL 2004*, (Tucson, Arizona; June 2004): 400.
6. J. Bekaert and H. Van de Sompel. "A Standards Based Solution for the Accurate Transfer of Digital Assets." *D-Lib Magazine* **11**(6) (June 2005).
7. OAICat.. http://www.oclc.org/research/software/oai/cat.htm

Personalized Digital E-library Service Using Users' Profile Information*

Wonik Park[1], Wonil Kim[2], Sanggil Kang[3], Hyunjin Lee[2], and Young-Kuk Kim[1]

[1] Department of Computer Engineering, Chungnam National University,
Daejeon, South Korea
{wonik78, ykim}@cnu.ac.kr
[2] College of Electronics and Information, Sejong University, Seoul, South Korea
wikim@sejong.ac.kr, enjoyaje@hotmail.com
[3] Department of Computer Science, University of Suwon, Gyeonggi-do, South Korea
sgkang@suwon.ac.kr

Abstract. We propose a personalized digital E-library system using a collaborative filtering technique, which provides a personalized search list according to users' preference. The proposed system analyzes the registered users' actions such as "clicking" and "borrowing" items. According to the different actions, we provide a weight for calculating the users' preference of each item. However, the list is uniformly provided to the individual users when they search with same keywords. In order to avoid the problem, we customize the order of items in the list according to whether there is any mismatching of profiles among registered users and target users or not.

1 Introduction

Most of the library search systems provide uniformly a list of items, such as books, papers, magazines, etc, to individual users without users' discretion. Usually users are reluctant to spend much time to look up the flood of the unwanted items in the list. To solve the problems, many researchers [1-3] developed the personalized digital library systems.

In this paper, we propose a personalized digital E-library system using a collaborative filtering technique, which provides a personalized search list according to users' preference. The proposed system analyzes the registered users' actions such as "clicking" and "borrowing" items. According to the different actions, we provide a weight for calculating the users' preference of each item. However, the list is uniformly provided to the individual users when they search with same keywords. In order to avoid the problem, we customize the order of items in the list according to whether there is any mismatching of profiles among registered users and target users or not.

The remainder of this paper consists as follows. Chapter 2 explains the proposed system in detail. Chapter 3 briefly explains the over architecture of our system. In Chapter 4, we show the simulated results of our system. Finally chapter 5 will conclude.

* This research was supported by the Ministry of Information and Communication, Korea, under the College Information Technology Research Center Support Program, grant number IITA-2005-C1090-0502-0016.

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 528 – 531, 2006.

2 Overall System Architecture

Fig. 1 shows the overall architecture of the proposed E-library service system. The architecture is composed of three modules such as Log Analyzer (LA), Personalization Inference (PI), and Dynamic HTML Generation Machine (DHGM). The LA module collects the registered users' clicking frequency of each item from their log information. If a target user requests a query, the PI module infers the personalized list by our algorithm explained in the following section, which utilizes the obtained users' clicking information, the borrowing information stored in the DB, and users' profile information stored in the Profile DB. The DHGM module shows the inferred personalized list to the target user through Displayed List.

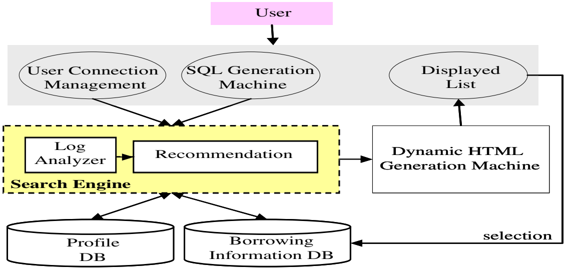


Fig. 1. The overall architecture of the proposed E-library service system

3 Proposed Digital E-library Service

For items retained in Digital E-library such as book, paper, magazine, etc, the preference of each item can be expressed by the actions taken by users. There are two types of actions such as "borrowing and "clicking" items. In conventional method, the preference of an item is expressed with the frequency of clicking for the item by users during a predetermined period, as seen in Equation (1).

$$\Pr ef_i = \sum_{k=1}^{K} c_{k,x_i} \tag{1}$$

where $Pref_i$ is the preference of item x_i, c_{k,x_i} is the frequency of clicking item x_i by user k and K is the total number of users registered in the Digital E-library. Equation (1) is the preference obtained using the action of clicking only. However, the action "borrowing" is usually stronger index for estimating the preference than the action clicking." In order to take into the consideration, we provide a weight w to the action "borrowing" in calculating the frequency of the action of clicking in Equation (1).

$$\Pr ef_i = \sum_{k=1}^{K} (c_{k,x_i} + w \cdot b_{k,x_i}) \tag{2}$$

where b_{k,x_i} is the frequency of borrowing item x_i by user k and $w > 1$. As seen in Equation (2), the accuracy of the preference depends on the value of w. As the frequency of borrowing an item increases, the preference for the item linearly increases with slop *w*. Based on the value of the preference, the order of items in the list as the result of a target user's query using one or more than one keyword is determined. However, the list is uniformly provided to the individual users when they search with same keywords. In order to avoid the problem, we customize the order of the list according to users' profile.

In general, the action information of users with the same profile as a target user is more useful for predicting the target user's usage behavior than that of users with different profile. For example, let a target user's profile information be u_t= (Student, Engineering, Computer Engineering, Junior). Also, there is two users' (User 1 and User 2) action information for the book C++ stored in the library database. User 1 with profile information u_1= (Student, Engineering, Computer Engineering, Junior) clicked the book C++ once. User 2 with profile information u_2= (Student, College of Art, Painting, Freshman) borrowed the book C++ once. In this case, even though the action of borrowing is more effective on computing the preference of the book C++ than the action of clicking, the action of clicking is more reliable because the profile information of User 1 is identical with the target user. In order to consider this problem for computing the preference, we modify the frequency of the actions by providing a penalty according to the degree of mismatching profiles between the users in the database and the target user.

$$\Pr ef_i = \sum_{k=1}^{K} p_k (c_{k,x_i} + w \cdot b_{k,x_i}) \tag{3}$$

where p_k is the penalty for the mismatching and $p_k \leq 1$. If there is no mismatching between user *k* and the target user then $p_k = 1$.

By using Equation (3) for computing the preference of each item, the personalized search list can be provided by according to different target users. Also, as seen in Equation (3), the accuracy of estimating the preference of each item depends on the values of the variables *w* and p_k. In the experimental section, we show the optimal values of those variables from the empirical experience.

4 Experiment

We implemented our system using the JAVA webserver in the Window NT environment. In the server, we used the JSDK which is Java servlet developer kit 1.4 to run our personalized E-library service program. MS SQL server 2000 was used as the relational database. Also, JDBC (Java Database Connectivity) was used in order to connect database with servlet.

Also, we made a sample E-library website with 1,000 book lists and collected the profile information and the actions of clicking and borrowing of 100 students registered in the Chungnam National University (CNU) in Korea and 50 faculties during one month from October 2005 to November 2005. Their actions were stored in the databases. For 10 target students and 10 target faculties whose actions are not stored in the database, we evaluated the performance of our system by comparing the degree of the satisfaction of our system with that of the E-library system provided by the

CNU. For the weight of borrowing action in Equation (3), we chose $w = 4$. Also, the penalties of mismatching between the k^{th} registered user and a target user are chosen as follows; $p_k = 0.6$. We evaluated the performances of the CNU E-library system and our personalized E-library system by surveying the satisfaction of both systems for each target user.

From Table. 1, 16 target users (3 for "very satisfactory", 13 for "satisfactory") out of 20 for our system expressed their satisfaction at our system, while only 2 target users at the CNU E-library system.

Table. 1. The result of the survey of the satisfaction of both systems for the 20 target users

Evaluation	Target users	
	CNU E-library	**Our system**
very satisfactory	0	**3**
satisfactory	2	**13**
dissatisfied	17	**3**
very dissatisfied	1	**1**

5 Conclusion

In this paper, we proposed the personalized E-library system by considering E-library users' actions such as "clicking" and "borrowing" and their profile information. From the experimental section, it is shown that our system can give satisfaction to E-library users, compared to the existing E-library system.

However, the weight of borrowing action and the penalties of mismatching used in the experiment were obtained from the exhaustive empirical experience. We need to do further study for developing an automatic algorithm in determining the weight and the penalties for various situations.

References

1. Bollacker, K.D., Lawrence S., Giles, C.L.: A System for Automatic Personalized Tracking of Scientific Literature on the Web. Proc. ACM Conference on Digital Libraries (1999) 105-113
2. Lee, W.P., Yang, T.H.: Personalizing Information Appliances: A Multi-agent Framework for TV Program Recommendations. Expert Systems with Applications, vol. 25, no. 3 (2003) 331-341
3. Kamba, T., Bharat, K., Albers, M.C.: An Interactive Personalized Newspaper on the Web. Proc. International World Wide Web Conference (1995) 159-170

Representing Aggregate Works in the Digital Library

George Buchanan[1], Jeremy Gow[2], Ann Blandford[2],
Jon Rimmer[3], and Claire Warwick[3]

[1] University of Wales, Swansea
g.r.buchanan@swansea.ac.uk
[2] UCL Interaction Centre, London
{j.gow, a.blandford}@ucl.ac.uk
[3] UCL SLAIS, London
{j.rimmer, c.warwick}@ucl.ac.uk

Abstract. This paper studies the challenge of representing aggregate works such as encyclopaedia, collected poems and journals in digital libraries. Reflecting on materials used by humanities academics, it demonstrates the complex range of aggregate types and the problems of representing this heterogeneity in the digital library interface. We demonstrate that aggregates are complex and pervasive, challenge many common assumptions and confuse the boundaries between organisational levels within the library. The challenge is amplified by concrete examples.

Keywords: Digital Libraries, Architecture, Collection Building.

1 Introduction

As more pre-digital humanities material is made available digitally, many collections now deal with aggregate works which associate a single identity with a set of atomic documents. But whilst these historic items are being digitised, historic forms of reference may be neglected. Locating an item within an aggregate requires searching and browsing to accurately reflect its structure.

One common and simple aggregate is the journal. If a collection is built of individual journal articles, then one document consistently represents one article, a journal issue is a set of articles, a volume a set of issues. It would appear logical that a similar approach should be effective for other aggregates. However, that is not the case. If a work is bound in two separate volumes, then it would make sense to separate between the two. However, that means that we now have two separate 'documents' in the library, which need to be linked for the purposes of browsing and searching. Counter-examples can also be found where multiple books are bound in one volume. An effective library will support retrieval under either criteria.

In addressing aggregate works, we presuppose the existence of an atomic 'document unit'. Aggregate works are defined as ordered trees with documents units at the leaves. This paper continues with an enumeration of aggregate features, followed with a review of problematic cases. We close with a discussion of related literature and the course for future research.

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 532–535, 2006.

2 Aggregate Structures in Practice

Here we enumerate some significant features of aggregate works. Note that these features are not all mutually exclusive:

Homogenous Aggregation. Each aggregated unit is of the same type.

Heterogenous Digital Forms. Though an aggregate work may be logically homogenous, its digital form may vary internally. e.g. digitisation occured over a period in which practice shifted.

Serial Aggregation. Aggregation from a series of related publications. e.g. journals or larger works that are published over many years.

Binding Aggregation. A work was printed and released as one item, but bound in separate volumes.

Composite Aggregation. When a work is published in parts, as with *serial aggregation*, but each part is itself bound within a different aggregate. e.g. 19th Century novels serialised in magazines.

Containing Aggregation. A work may be small and unavailable in its own right, but available contained within larger works which are not themselves aggregates, e.g. a poem within a work of fiction.

Heterogenous Aggregation. A work is created from units of diverse types. For instance, newspapers and journals contain articles of different types that may need to be distinguished in the DL interface.

Supplementary Aggregation. Where an original work is supplemented by further material, possibly by another author.

Incomplete Aggregation. Some aggregates are incomplete, either because they were not fully published or because a collection is only partial.

Variable Aggregation. Different versions of an aggregate work may bring together different material, or different versions of the same material.

Furthermore, the boundary between external and internal document structure is not fixed, and many of the issues above may also occur *within* a document. What is important, from the view of a DL system, is that the treatment of internal and external aggregation are treated consistently in the DL architecture and also in the user interface, to ease the task of readers and librarians alike.

3 Difficult Cases

Our own experience on realising aggregates is based on the Greenstone DL system [8], and DSpace [6]. Simple cases such as journal collections result in few problems. However, beyond such regular structures, problems rapidly multiply. In a collection of literature the scale of items varies from a short stories to a multi-volume "epics". If we faithfully replicate the physical text, some items will be multi-volume, whilst a single volume may contain several works. The concept of 'volume' thus becomes problematic.

Indexing a collection by volume conflates works that share the same volume, whilst indexing by works only will conflate volumes of the same work. Clearly, neither solution is optimal: the natural conclusion is to index by the smallest unit (work, volume) and aggregate upwards to unify elements of the same item. This underlying storage can be represented in different ways in the library interface: e.g., matches against a single search for separate volumes of the same work can be unified in the search result list. This option is already available in Greenstone [8], and can be achieved in DSpace with careful configuration and effort.

During browsing, however, the contradictory use of volume (as a part or as an aggregate) will still emerge in some form or other. One can distinguish the part-of and aggregate-of styles of volume by introducing a three-level hierarchy and using disciminating labels for the top and bottom levels. Many items are represented by only one item at each level, and as reported in [7] such simple single-child relationships should be pruned so that unnecessary interaction is minimised. Thus to improve the interactional efficiency, the experienced hierarchy becomes irregular. The issues of unifying hierarchy nodes in search result lists remains a problem (though this can be achieved in Greenstone).

We now focus on complex cases with increasing degrees of difficulty, particularly *containing aggregation* and *composite aggregation.* Composite aggregates represent particularly problematic structures. Serialised fiction such as Conan Doyles *Gang of Four* disrupts DL assumptions in its original form. If each newspaper in a collection is stored as a single document, then a reader will need to map the Gang of Four in its original context to particular editions of the correct publication – which they may not know!

An alternative approach would be to extract and record the elements of the story as one DL document, tidily avoiding the problem for a searcher specifically looking for the Gang of Four, but conversely divorcing it from its original context - to connect each article with its context in the original magazine, the user must in fact engage in the 'hunting' of articles we apparently just avoided. Such contextual interpretation is the knub of many items of humanities research. Clearly, an optimal approach allows both the recovery of the original composited piece, and the magazines of which it was part.

4 Related Work

The difficulties of the representation of aggregate works in digital libraries has already received attention: e.g. Hickey and O'Neill [1] note problems encountered in applying FRBR [3]. O'Neill proposes treating aggregates as published volumes of more than one work, and to avoid recording aggregates as works in their own right. This introduces an inconsistency with the accepted FRBR model where every published volume (*manifestation*) is an instance of a single work.

Two electronic document standards support aggregate works: TEI [4] and METS [2]. In both cases, aggregates are achieved by pointers to their parts, to create a whole. TEI primarily uses pointers between parts of the aggregate, whereas in METS a central document contains references to part or whole other METS documents. Aggregates have been poorly represented in DL systems: e.g. DSpace [6] and Greenstone [8] focus on treating collections as sets of objects, with a hierarchical classification structure. Aggregates can be represented using the classification structure, but at the loss of consistent treatment of aggregates across both searching and browsing. In library science, the need to find and recover texts via bound volumes has emphasised the same approaches we see in DL systems. Aggregates are generally indexed by part where the parts are discrete works: e.g. the British Library binds brief tracts together in volumes, but each tract in a volume is indexed separately. Conversely, multi-volume works are usually, but not universally, indexed by one entry.

Svenonius [5], p. 103, notes that there are two potential routes to relating aggregates with their constituent parts: first, formal linkage structures; second, providing descriptive aggregation (meta–)data for each item. The latter approach, though informal and easy to apply, leaves much of the retrieval work with the user, and greater room for mismatches between the descriptive data and the corresponding description of the part or aggregate in the catalogue index.

5 Conclusion

We described above a number of different forms of aggregate work found in the humanities. Simple forms may be supported in DLs with only small shortcomings in representation. However, more complex forms of aggregation which occur frequently in historic literature map less readily to existing DL architectures and interfaces. In our research, we wish to investigate further the appropriate interactions to support the occurrence of aggregates in search result lists, and the location of desired aggregates in the course of information seeking.

Acknowledgements. This research is supported by EPSRC Grant GR/S84798.

References

1. T. B. Hickey and E. T. O'Neill. Frbrizing oclc's worldcat. *Cataloging and Classification Quarterly*, 39:239–251, 2005.
2. Library of Congress. *Metadata Encoding and Transmission Standard (METS).*
3. S. G. on the Functional Requirements for Bibliographic Records. *Functional requirements for bibliographic records.* K.G. Saur, 1998.
4. C. Sperberg-McQueen and L. Burnard, editors. *Guidelines for Electronic Text Encoding and Interchange.* TEI P3 Text Encoding Initiative, Oxford, 1999.
5. E. Svenonius. *The Intellectual Foundation of Information Organization.* Digital Libraries and Electronic Publishing. MIT Press, 2000.
6. R. Tansley, M. Smith, and J. H. Walker. The dspace open source digital asset management system: Challenges and opportunities. In *Procs. European Conference on Digital Libraries*, pages 242–253. Springer, 2005.
7. Y. L. Theng, E. Duncker, N. Mohd-Nasir, G. Buchanan, and H. Thimbleby. Design guidelines and user-centred digital libraries. In *Proc. 3rd European Conf. for Digital Libraries, ECDL*, pages 125–134. Springer-Verlag, 1999.
8. I. H. Witten, S. J. Boddie, D. Bainbridge, and R. J. McNab. Greenstone: a comprehensive open-source digital library software system. In *Proc. ACM conference on Digital libraries*, pages 113–121. ACM Press, 2000.

Scientific Evaluation of a DLMS: A Service for Evaluating Information Access Components

Giorgio Maria Di Nunzio and Nicola Ferro

Department of Information Engineering – University of Padua
Via Gradenigo, 6/b – 35131 Padova – Italy
{dinunzio, ferro}@dei.unipd.it

Abstract. In this paper, we propose an architecture for a service able to manage, enrich, and support the interpretation of the scientific data produced during the evaluation of information access and extraction components of a *Digital Library Management System (DLMS)*. Moreover, we describe a first prototype, which implements the proposed service.

1 Introduction

As observed in [4], "*Digital Library (DL)* development must move *from an art to a science*" in order to design and develop *Digital Library Management Systems (DLMSs)*, based on reliable and extensible services. This shift from DLs to service-based DLMSs and the requirement for improved reliability points out, among other issues, the need of proper evaluation methodologies in order to assess a DLMS along different dimensions.

The evaluation itself of a DLMS turns out to be a scientific activity whose outcomes, such as performance analyses and measurements, constitute a kind of *scientific data* that need to be properly considered and used for the design and development of DLMS components and services. These scientific data should, in turn, be managed by a DLMS which takes care of supporting their enrichment and interpretation. We propose to name this type of DLMS a *scientific reflection DLMS*, since it deals with scientific data, information, and interpretations about the design and development of another DLMS.

There are many aspects to take into consideration when evaluating a DLMS, such as information access capabilities, interaction with users, and so on. As a consequence, the scientific reflection DLMS should be constituted by different and cooperating services, each one focused on supporting the evaluation of one of the aspects mentioned above. In particular, we face the problem of designing and developing a service for evaluating the information access components of a DLMS.

2 Design of an Information Access Evaluation Service for a Scientific Reflection DLMS

Figure 1 shows the architecture of the proposed service. It consists of three layers – data, application and interface logic layers – in order to achieve a better

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 536–539, 2006.

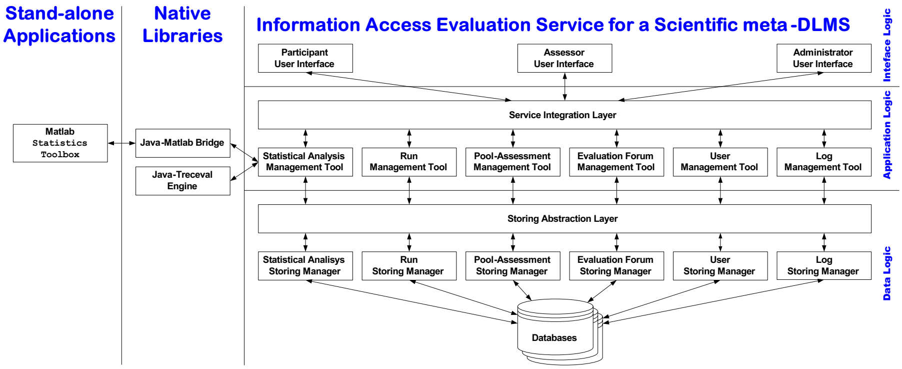

Stand-alone Applications
Native Libraries
Information Access Evaluation Service for a Scientific meta -DLMS
Inteface Logic
Application Logic
Data Logic
Participant User Interface
Assessor User Interface
Administrator User Interface
Service Integration Layer
Matlab Statistics Toolbox
Java-Matlab Bridge
Java-Treceval Engine
Statistical Analysis Management Tool
Run Management Tool
Pool-Assessment Management Tool
Evaluation Forum Management Tool
User Management Tool
Log Management Tool
Storing Abstraction Layer
Statistical Analisys Storing Manager
Run Storing Manager
Pool-Assessment Storing Manager
Evaluation Forum Storing Manager
User Storing Manager
Log Storing Manager
Databases

Fig. 1. Architecture of a service for supporting the evaluation of the information access components of a DLMS

modularity and to properly describe the behavior of the service by isolating specific functionalities at the proper layer.

Data Logic. The data logic layer deals with the persistence of the different information objects coming from the upper layers. There is a set of "storing managers" dedicated to storing the submitted experiments, the relevance assessments and so on. We adopt the *Data Access Object (DAO)*[1] and the *Transfer Object (TO)*[1] design patterns. The DAO implements the access mechanism required to work with the underlying data source, acting as an adapter between the upper layers and the data source.

In addition to the other storing managers, there is the *log storing manager* which fine traces both system and user events. It captures information such as the user name, the *Internet Protocol (IP)* address of the connecting host, the action that has been invoked by the user, any error condition, and so on. Thus, besides offering us a log of the system and user activities, the log storing manager allows us to fine trace the provenance of each piece of data from its entrance in the system to every further processing on it.

Finally, on top of the various "storing managers" there is the *Storing Abstraction Layer (SAL)* which hides the details about the storage management to the upper layers. In this way, the addition of a new "storing manager" is totally transparent for the upper layers.

Application Logic. The application logic layer deals with the flow of operations within *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)*. It provides a set of tools capable of managing high-level tasks, such as experiment submission, pool assessment, statistical analysis of an experiment.

For example, the *Statistical Analysis Management Tool (SAMT)* offers the functionalities needed to conduct a statistical analysis on a set of experiments.

[1] `http://java.sun.com/blueprints/corej2eepatterns/Patterns/`

In order to ensure comparability and reliability, the SAMT makes uses of well-known and widely used tools to implement the statistical tests, so that everyone can replicate the same test, even if he has no access to the service. In the architecture, the MATLAB Statistics Toolbox[2] has been adopted, and an additional library has been implemented to allow our service to access MATLAB in a programmatic way. As an additional example aimed at wide comparability and acceptance of the tools, a further library provides an interface for our service towards the `trec_eval` package[3], which represents the standard tool for computing the basic performance figures, such as precision and recall.

Finally, the *Service Integration Layer (SIL)* provides the interface logic layer with a uniform and integrated access to the various tools. As we noticed in the case of the SAL, thanks to the SIL also the addition of new tools is transparent for the interface logic layer.

Interface Logic. It is the highest level of the architecture, and it is the access point for the user to interact with the system. It provides specialised *User Interfaces (UIs)* for different types of users, that are the participants, the assessors, and the administrators. Note that, thanks to the abstraction provided by the application logic layer, different kind of UIs can be provided, either stand-alone applications or Web-based applications.

3 The First Running Prototype

The proposed service has been implemented in a first prototype, called *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)* [2], and it has been tested in the context of the *Cross-Language Evaluation Forum (CLEF)* 2005 evaluation campaign. The initial prototype moves a first step in the direction of an information access evaluation service for scientific reflection DLMSs, by providing support for: the management of an evaluation forum: the track set-up, the harvesting of documents, the management of the subscription of participants to tracks; the management of submission of experiments, the collection of metadata about experiments, and their validation; the creation of document pools and the management of relevance assessment; common statistical analysis tools for allowing the comparison of the experiments; common tools for summarizing, producing reports and graphs on the measured performances and conducted analyses.

The prototype was successfully adopted during the CLEF 2005 campaign. It was used by nearly 30 participants spread over 15 different nations, who submitted more than 530 experiments; then 15 assessors assessed more than 160,000 documents in seven different languages, including Russian and Bulgarian which do not have a latin alphabet. It was then used for producing reports and overview graphs about the submitted experiments [1,3].

[2] `http://www.mathworks.com/products/statistics/`

[3] `ftp://ftp.cs.cornell.edu/pub/smart/`

DIRECT has been developed by using the Java[4] programming language. We used the PostgreSQL[5] *DataBase Management System (DBMS)* for performing the actual storage of the data. Finally, a Web-based interface, which make the service easily accessible to end-users without the need of installing any kind of software, has been developed by using the Apache STRUTS[6] framework.

4 Conclusions

We proposed the architecture of a service which evaluates the information access components of a DLMS. This service should be part of a wider *scientific reflection DLMS*, which allows for enriching and interpreting the scientific data produced during the evaluation of a DLMS. A first prototype of the proposed service has been implemented and widely tested during the CLEF 2005 evaluation campaign.

Acknowledgements

The authors would like to thank Carol Peters, coordinator of CLEF, for fruitful discussions in the context of the CLEF effort. The work reported in this paper has been partially supported by the DELOS Network of Excellence on Digital Libraries, as part of the Information Society Technologies (IST) Program of the European Commission (Contract G038-507618).

References

1. G. M. Di Nunzio and N. Ferro. Appendix A. Results of the Core Tracks and Domain-Specific Tracks. In *Working Notes for the CLEF 2005 Workshop*. `http://www.clef-campaign.org/2005/working_notes/workingnotes2005/appendix_a.pdf`, 2005.
2. G. M. Di Nunzio and N. Ferro. DIRECT: a System for Evaluating Information Access Components of Digital Libraries. In *Proc. 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005)*, pages 483–484. LNCS 3652, Springer, Heidelberg, Germany, 2005.
3. G. M. Di Nunzio, N. Ferro, G. J. F. Jones, and C. Peters. CLEF 2005: Ad Hoc Track Overview. In *Accessing Multilingual Information Repositories: Sixth Workshop of the Cross–Language Evaluation Forum (CLEF 2005). Revised Selected Papers*. LNCS 4022, Springer, Heidelberg, Germany (in print), 2006.
4. Y. Ioannidis, D. Maier, S. Abiteboul, P. Buneman, S. Davidson, E. A. Fox, A. Halevy, C. Knoblock, F. Rabitti, H.-J. Schek, and G. Weikum. Digital library information-technology infrastructures. *International Journal on Digital Libraries*, 5(4):266–274, 2005.

[4] `http://java.sun.com/`

[5] `http://www.postgresql.org/`

[6] `http://struts.apache.org/`

SIERRA – A Superimposed Application for Enhanced Image Description and Retrieval

Uma Murthy[1], Ricardo da S. Torres[2], and Edward A. Fox[1]

[1] Department of Computer Science, Virginia Tech,
Blacksburg, VA 24061, USA
{umurthy, fox}@vt.edu
[2] Institute of Computing, State University of Campinas,
Campinas, SP, Brazil, 13084-851
rtorres@ic.unicamp.br

Abstract. In this demo proposal, we describe our prototype application, SIERRA, which combines text-based and content-based image retrieval and allows users to link together image content of varying document granularity with related data like annotations. To achieve this, we use the concept of superimposed information (SI), which enables users to (a) deal with information of varying granularity (sub-document to complete document), and (b) select or work with information elements at sub-document level while retaining the original context.

1 Description

In many image-based applications, like biomedical teaching, research, and diagnosis, there is need to link (or integrate) image content with other multimedia information: text annotations, metadata (keywords or ontological terms), audio-visual presentations, etc. Not only does this contribute to richer image descriptions, it also helps in more effective retrieval of images and related information [10]. Further, for complex images (e.g., images with plenty of detail, or with specific hard-to find details), there may be a need to isolate and work with parts of images (meaningful objects within the image) without losing the original context (the actual image). For example, an ichthyologist may want to annotate a particular part of a fish after seeing annotations from other ichthyologists on the same type of fish. Yet, current image-based systems either focus on content-based [8] or text-based descriptions [3, 4]. Some systems [1, 2, 6, 7, 12], which combine both techniques to enhance the image annotation process, provide limited support for linking image content, at varying document granularity, to other multimedia content (like ontological terms, video descriptions, etc.).

We have developed SIERRA, an application which combines text- and content-based image retrieval so users can relate images, at varying document granularity, to other multimedia content, applying the concept of superimposed information (SI). SI refers to new information (or new interpretations) laid over existing information [9] (like bookmarks, annotations, etc.). Superimposed applications (SAs) allow users to lay new interpretations over existing or base information. SAs employ “marks”, which are references to selected regions within base information [11]. SAs enable

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 540 – 543, 2006.

users to (a) deal with information of varying granularity, and (b) select or work with information elements at sub-document level while retaining the original context.

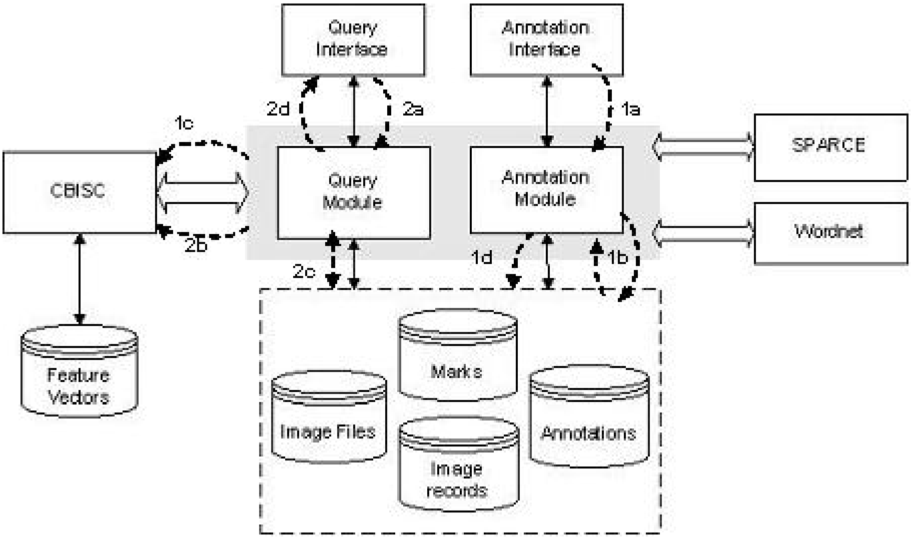
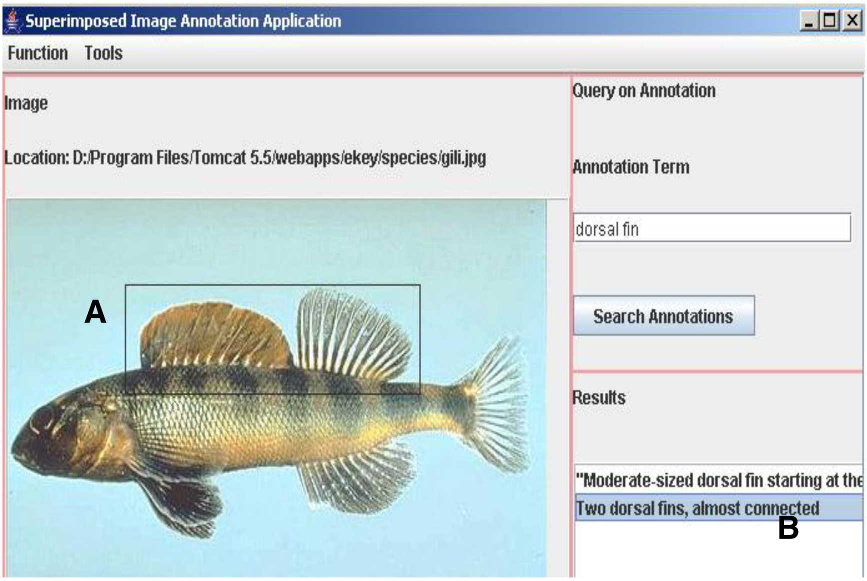


Fig. 1. The high-level design of SIERRA and labeled steps involved in two major functions of SIERRA – 1) the marking and annotation process, and 2) the retrieval process

Figure 1 shows the high-level design of SIERRA. It consists of two main modules – the annotation module and the query module. The design is such that other existing modules may be plugged in to facilitate richer image description and retrieval. SIERRA makes use of the Content-Based Image Search Component (CBISC) [13], an OAI-compliant component that supports queries on image collections. It retrieves images similar to a user-defined pattern (e.g., color layout of an image, image sketch, etc.) based on content properties (e.g., shape, color, or texture), which are often encoded in terms of image descriptors. In addition, we foresee integration with other types of components like the ontology WordNet [5] (for suggesting annotation terms), and the Superimposed Pluggable Architecture for Contexts and Excerpts (SPARCE) – middleware for managing "marks" over text, audio, and video content [11]. Integration with SPARCE will enable associating image marks and annotations with marks in other content types.

Two major applications of SIERRA include: 1) the image marking and annotation process, and 2) the image/mark/annotation retrieval process. Figure 1 traces the high-level steps of scenarios involving each of these processes. 1a) the user identifies an image, marks a region of interest and annotates that region with keywords; 1b) a mark is created and all mark-relevant information is stored; 1c) the content of the sub-image referred to by the mark is stored in the CBISC; 1d) annotation information is stored; 2a) the user identifies an image, then marks (selects) a region within the image and uses this mark to query SIERRA; 2b) SIERRA uses the sub-image referenced by the mark created by the user to query the CBISC and get a list of images or marks similar to the queried mark. 2c) all annotations associated with the result images/marks are retrieved; 2d) the user is able to view the result images/marks with associated annotations.



Location: D:/Program Files/Tomcat 5.5/webapps/ekey/species/gili.jpg

Annotation Term: dorsal fin

"Moderate-sized dorsal fin starting at the

Two dorsal fins, almost connected

Fig. 2. A snapshot of the initial prototype of SIERRA. All annotations with the phrase "dorsal fin" are listed. On selecting an annotation, tmhe associated mark (and the containing image) is displayed. A) Mark associated with selected annotation; B) Selected annotation.

The current prototype of SIERRA (see Figure 2) allows users to select parts of images and associate them with text annotations. Then, users can retrieve information as annotations and associated marks in two ways, either for (1) a specified image, or (2) annotations containing specified query terms. The first capability illustrates how this SA differs from a typical hypermedia application, in that important work can be done just with the marks, ignoring the base information.

This prototype has been developed in Java and makes use of the Java 2D API for image manipulation. Data is stored in a PostgreSQL database.

We are integrating this prototype with our content-based image search component [13] to extract content from complete images and marks. We will then undertake a formative usability evaluation on the prototype. Future work on this application includes integration with the ontology WordNet [5] and with SPARCE [11].

Acknowledgements

This work is funded in part by NSF DUE-0435059, under the NSDL research track. Ricardo Torres' work is funded in part by CAPES, FAPESP, FAEPEX, the CNPq WebMaps and AgroFlow project, and by a Microsoft eScience grant.

References

1. Barnard, K., Duygulu, P., Freitas, N.d., Forsyth, D., Blei, D. and Jordan, M.I. Matching Words and Pictures. *Journal of Machine Learning Research*, *3* (6): 1107-1135.
2. Benjamin, B.B., PhotoMesa: a zoomable image browser using quantum treemaps and bubblemaps. In *Proceedings of the 14th annual ACM symposium on user interface software and technology*, Orlando, Florida, 2001, ACM Press, 71 - 80.
3. Chen-Yu, L., Von-Wun, S. and Yi-Ting, F., How to annotate an image? the need of an image annotation guide agent. In *Proceedings of the 4th ACM/IEEE-CS joint conference on digital libraries*, Tuscon, AZ, USA, 2004, ACM Press, 394 - 394.
4. Elin, G., Rohlfing, M. and Parenti, M. Fotonotes.net - Image Annotation Standard and Scripts, 2004, http://www.fotonotes.net/.
5. Fellbaum, C. (ed.), *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA, 2001.
6. Freitas, R. and Torres, R., OntoSAIA: Um Ambiente Baseado em Ontologias para Recuperação e Anotação Semi-Automática de Imagens. In *Proceedings of Workshop em Bibliotecas Digitais*, Uberlândia, MG, Brasil, 2005, 60-79.
7. Halaschek-Wiener, C., Schain, A., Golbeck, J., Grove, M., Parsia, B. and Hendler, J., A Flexible Approach for Managing Digital Images on the Semantic Web. *Presented at* the 5th International Workshop on Knowledge Markup and Semantic Annotation, Galway, Ireland, 2005.
8. Lieberman, H., Rosenzweig, E. and Singh, P. Aria: an agent for annotating and retrieving images. *Computer*, *34* (7): 57-62.
9. Maier, D. and Delcambre, L., Superimposed Information for the Internet. In *Proceedings of the WebDB Workshop*, Philadelphia, PA, USA, 1999, 1-9.
10. Muller, H., Michoux, N., Bandon, D. and Geissbuhler, A. A review of content-based image retrieval systems in medical applications - clinical benefits and future directions. *International Journal of Medical Informatics*, *73* (1): 1-23.
11. Murthy, S., Maier, D., Delcambre, L. and Bowers, S., Putting Integrated Information into Context: Superimposing Conceptual Models with SPARCE. In *Proceedings of the First Asia-Pacific Conference of Conceptual Modeling*, Denedin, New Zealand, 2004, 71-80.
12. Stein, A. COLLATE - Collaboratory for Annotation, Indexing and Retrieval of Digitized Historical Archive Material. IST-1999-20882, Fraunhofer IPSI, Dolivostrasse 15, D-64293, Place Published, 2004,
13. Torres, R., Medeiros, C.B., Goncalves, M.A. and Fox, E.A. A Digital Library Framework for Biodiversity Information System. *International Journal on Digital Libraries*, *6* (1): 3 - 17.

The Nautical Archaeology Digital Library

Carlos Monroy[1], Nicholas Parks[1], Richard Furuta[1], and Filipe Castro[2]

[1] Center for the Study of Digital Libraries, Texas A&M University,
College Station, TX 77843-3112, USA
{cmonroy, parksnj, furuta}@csdl.tamu.edu
[2] Institute of Nautical Archaeology, Texas A&M University,
College Station, TX 77843-4352, USA
fvcastro@tamu.edu

Abstract. In Nautical Archaeology, the study of components and objects creates a complex environment for scholars and researchers. Nautical archaeologists access, manipulate, study, and consult a variety of sources from different media, geographical origins, ages, and languages. Representing underwater excavations is a challenging endeavor due to the large amount of information and data in heterogeneous media and sources that must be structured, segmented, categorized, indexed, and integrated. We are creating a Nautical Archaeology Digital Library that will a) efficiently catalog, store, and manage artifacts and ship remains along with associated information from underwater archeological excavations, b) integrate heterogeneous data sources in different media to facilitate research work, c) incorporate historic sources to help in the study of current artifacts, d) provide visualization tools to help researchers manipulate, observe, study, and analyze artifacts and their relationships; and e) incorporate algorithm and visualization based mechanisms for ship reconstruction.

1 Introduction

The research methodology in Nautical Archaeology is evidence-based—data and artifacts gathered from the field provide the basis for evaluation of hypothesized relationships. The range of sources brought to bear in evaluating hypotheses is wide-ranging both in scope and also in time—sources range from hundreds-of-years-old historical treatises to digitized video indexed to data streams from modern-day satellite-based global positioning systems. In addition, data-gathering—i.e., surveying a site—is, by its nature, destructive, so the ability to validate a project's findings independently rests on the availability and completeness of the data and metadata obtained during the site survey. Clearly, relationships among these sources are complex.

We are developing a digital library framework for Nautical Archaeology that will provide a) flexible cross-linking of heterogeneous content in a dynamically-growing collection, b) flexible use of annotations to enhance community access while respecting individual information rights, c) incorporation and management of uncertain data, d) digital library replication and synchronization, and e) general applications of visualizations based on 2D grids. In our poster we will cover five major research areas in the context of Nautical Archaeology pertaining to: the excavation site, the archaeological recovery process, the artifacts collection, shipbuilding treatises, and ship modeling and reconstruction. Each area will include their corresponding source materials.

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 544 – 547, 2006.

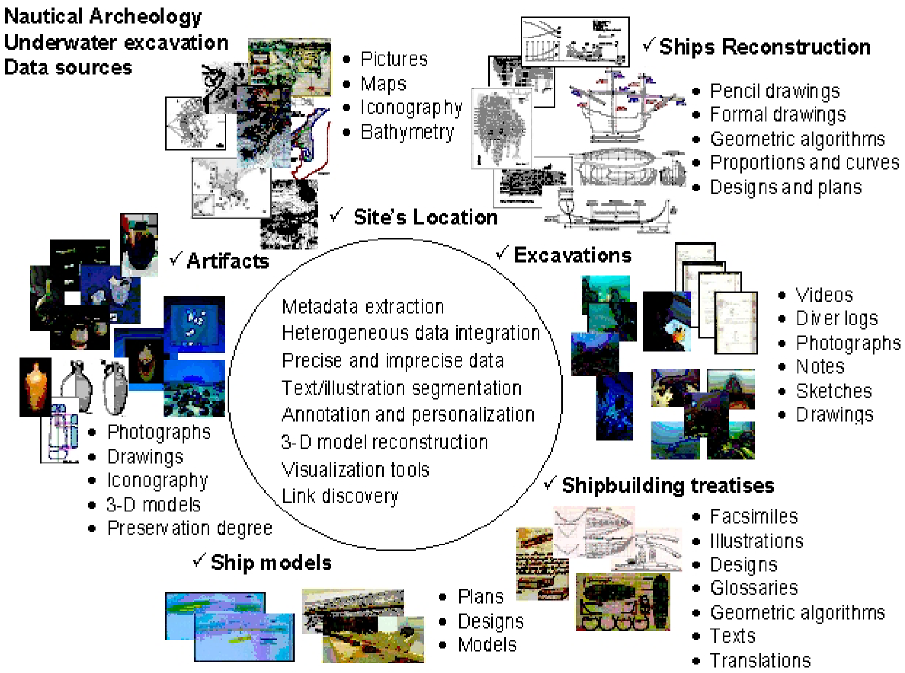


Fig. 1. The Nautical Archaeology Digital Library's data sources. The extent of the data sources involved is illustrated by the elements enumerated outside the central circle. Tasks to be addressed in the project are shown inside the circle.

The associated tasks to the aforementioned areas can be grouped as follows: a) developing a model for mapping an underwater archeological excavation site, b) establishing a protocol for storing, managing, and organizing information related to a shipwreck, c) creating a framework to enable the integration of heterogeneous data sources and media, d) developing new ways for structuring and accessing ancient shipbuilding treatises, and e) providing computational assistance for the identification and placement of ship fragments to allow ship reconstruction.

Figure 1 illustrates the scope of the proposed digital library, each area lists the data sources required. Tasks to be performed are listed in the circle at the center of the illustration.

Our project's philosophy is to investigate extensions to other's terrestrial archaeological digital libraries with the goal of addressing the unique characteristics of nautical archaeology. The result will be a resource of value to scholars and of interest to the general public.

2 Current NADL Project Activities

Our approach in the creation of this digital library is to focus not only on the collection itself but also on the work practices of the primary users. Thus, we are developing a suite of tools to assist nautical archaeologists in their scholarly work. The major goal is to support the complete archaeological process from site discovery and excavation to conservation and publication. Therefore, we must also support the archeologist in both the connected and connectionless work environments; in essence, we propose a mobile ubiquitous digital library system.

To make a successful ubiquitous system for humanities practitioners, one needs to understand the work practices of the scholars involved. In this context, the NADL development will employ ethnographic methods in addition to traditional software usability methods. To date, we have hours of audio interviews and DVDs of captured observations of the archeologist at work; which will help us develop a tool kit that users will actually use. Presently we have already fielded a prototype tool, code-named "OnScene," to cope with the tasks performed at the excavation site. The prototype system will be used this summer on site in Portugal, and the experience gained will guide subsequent refinements.

Archaeological excavations have a one-time component; there exists but one chance of discovery, one instance to capture material in spatial context, and there is only one chance to perform point of capture data collection. This mobile connectionless environment creates the most data points and currently is where much information is lost. Additionally, incomplete, inaccurate, and subjective information makes the information technology needs of nautical archaeologists unique. In other domains, information technology tools deal with the distribution, organization, and understanding of content. Here, the focus is trying to assist the archeologist understand the information available without the archeologist being aware of a digital library.

Further, a field excavation typically generates thousands of artifacts and other data points in a short period of time. The content types are far ranging from simple spread sheets, images, and video to the content generated by specialized archeological software. In fact, the fieldwork of an archeologist represents only a fraction of the time dedicated in an investigation. Archaeologists spend more time trying to understand the material recovered using methods familiar to other investigative scientists. Thus, our digital repository will automatically generate associations among recovered material. Also, we will assist those that use this tool kit and repository to create associations themselves, similar to tagging.

Moreover, it is important to understand that archeological excavations are multi-year (if not multi-decade) endeavors where the investigators may and do change. In many instances also, objects recovered are shrouded in a concretion. Thus, these objects tend to spend years as unknown entities while they endure the conservation process that is time and work intensive. Such changes will result in a discontinuity in the pace of the excavation as well as the focus.

The NADL team has been focusing on designing the product architecture to support our goals. As much as possible, we seek to build from existing projects, both in the archaeological domain and also in other digital libraries activities. Projects within the archaeological domain that are influencing our initial thoughts include ETANA [2], which has developed a system that handles dissimilar content dissemination through OAI-MHP [3], and the work of the Alexandria Institute, which has successfully articulated the need for adaptive representation of materials through their ArchaeoML work [1].

Our application domain of Nautical Archaeology raises an interesting set of problems. Reconstructing composite objects—such as ships—from incomplete or damaged sources requires, among other aids, the combination of algorithmic techniques and visualization tools. Ship reconstruction requires intensive querying of large amounts of dynamic information as new timbers are recovered and added into the repository. Once triaged, further visualization of the fragment in the context of a whole can help the researcher evaluate the suggested alternatives.

The repository includes timbers from other excavations as well as knowledge of the ideal characteristics, encoded from shipbuilding treatises. Shipbuilding treatises are technical manuals that include text describing the shipbuilding process, ship's proportions, illustrations of the pieces and components, and assembling instructions. The relevant treatises come from a variety of countries, kingdoms, and empires, over a span of several centuries, mainly between the 15th and 19th centuries. A challenging aspect in establishing relationships between treatises with ship remains is the capability to map them together. Further, representations and models of these relationships are important in establishing and/or validating hypotheses about the components of a ship, the construction techniques and geometric algorithms used, and the building sequences followed.

The NADL activity is still in its initial stages. Its progress can be followed at http://nadl.tamu.edu/.

Acknowledgements

This material is based on work supported by the National Science Foundation under Grant No. IIS-0534314.

References

1. Schloen, D.: Online Cultural Heritage Research Environment. The Oriental Institute of the University of Chicago http://ochre.lib.uchicago.edu/index.htm
2. Shen, R., Gonçalves, M., Fan, W., and Fox, E.: Requirements Gathering and Modeling of Domain-Specific Digital Libraries with the 5S Framework: An Archaeological Case Study with ETANA. In Proceedings ECDL2005, Vienna, 2005.
3. Van de Sompel, H., Nelson, M., Lagoze, C., Warner, S.,: Resource Harvesting within the OAI-PMH Framework. D-Lib Magazine, December 2004.

The SINAMED and ISIS Projects: Applying Text Mining Techniques to Improve Access to a Medical Digital Library

Manuel de Buenaga[1], Manuel Maña[2], Diego Gachet[1], and Jacinto Mata[2]

[1] Universidad Europea de Madrid – Escuela Superior Politécnica
28670 Villaviciosa de Odón, Madrid, España
{buenaga, gachet}@uem.es
[2] Universidad de Huelva – Dpto. Ing. Electrón., Sistemas Informáticos y Aut.
Escuela Politécnica Superior
21819 Palos de la Frontera, Huelva, España
manuel.mana@diesia.uhu.es, mata@uhu.es

Abstract. Intelligent information access systems integrate text mining and content analysis capabilities as a relevant element in an increasing way. In this paper we present our work focused on the integration of text categorization and summarization to improve information access on a specific medical domain, patient clinical records and related scientific documentation, in the framework of two different research projects: SINAMED and ISIS, developed by a consortium of two research groups from two universities, one hospital and one software development firm. SINAMED has a basic research orientation and its goal is to design new text categorization and summarization algorithms based on the utilization of lexical resources in the biomedical domain. ISIS is a R&D project with a more applied and technology-transfer orientation, focused on more direct practical aspects of the utilization in a concrete public health institution.

1 Project Goals

The SINAMED and ISIS projects are focused on information access on a specific biomedical domain: patient clinical records and related scientific documentation. These two projects have a strong interrelation and also different and complementary orientation. The SINAMED[1] project has a main orientation of basic research, focused on the design and the integration of automatic text summarization and categorization algorithms to improve access to bilingual information in the biomedical domain. The ISIS[2] project has a more applied and technology-transfer orientation, and its aim is the improvement in the intelligent access to the medical information, having in mind

[1] The research described in this paper has been partially supported by the Spanish Ministry of Education and Science and the European Union from the European Regional Development Fund (ERDF) - (TIN2005-08988-C02-01 and TIN2005-08988-C02-02).

[2] The research described in this paper has been partially supported by the Spanish Ministry of Industry and the European Union from the European Regional Development Fund (ERDF) - (FIT-350200-2005-16).

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 548 – 551, 2006.

doctors and patients as end users. It is focused on providing advanced and more effective tools than the current ones for the search, localization, use, and understanding of different sources of medical information.

2 The Medical Domain

The medical information is voluminous, heterogeneous and of extreme complexity. One of the factors with a major repercussion in the heterogeneity of the medical content is the source diversity. Each source (scientific papers, databases of summaries, structured or semi-structured databases, Web services or clinical records of patients) has several features. For example, the existence or not existence of an external structure for the document, the occurrence of free text together with structured data (tables with clinical results) or the length of the documents. These differences in domain, structure and scale hinder the development of robust and independent systems that facilitate the access to this kind of content.

Medical Documentation: Considering, for instance, the scientific medical articles, there are thousands of scientific journals in English language, and the problem grows if we consider other languages and other sources. Medline, the most important and consulted bibliographical database in the biomedical domain, constitutes a main example. Medline contains more than 13 million references, with an increment between 1.500 and 3.500 references per day. This huge volume of articles makes the experts difficult to take advantage of the whole published and interested information.

The Patient Clinical Record: The patient clinical record is defined as the set of documents (data, assessments and other type of information) that are generated throughout the assistance process of a patient. The system of clinical record sheets presents many drawbacks (unreadable information, chaos, absence of consistency, questionable availability, uncertain confidentiality guarantee, damage in the documents,…) that could largely be corrected with the usage of electronic clinical records. Some of the advantages of the electronic clinical record are: a better accessibility to the information and an improvement in the confidentiality; data homogenization; prescription filled in an automatic way; overall view of the patient; coordination of medical treatments; gathering of the whole information of a patient.

The combination of a scientific information system with the electronic clinical record would help doctors to make decisions, to decrease the mistakes and the clinical variability and to increase the patient's safety.

3 Text Mining Techniques

In the projects that we present in this paper, we propose to integrate text categorization and summarization techniques into the searching and browsing processes. We expect that a better organization could help users to feel less overwhelmed by the amount of information and to get a better understanding of the information available in the retrieved documents [1, 2].

Text Categorization: Automated text categorization can be applied, for example, to catalogue medical reports using standards descriptors, as the Medical Subject Headings (MeSH). However, the language variability and the lack of the needed data for an effective learning limits the effectiveness of these systems. Also, text categorization has rarely been applied in biomedical environment [3, 4] and the use of this technique on medical information writing in Spanish is virtually nonexistent.

The mentioned problems can be dealt with the use of lexical semantic resources. The techniques presented in these works are specially applicable to the medical information, since there are available specific resources as the Unified Medical Language System (UMLS).

Text Summarization: In information access environments, summaries (single-document or multi-document) have proved its utility, improving the effectiveness of several tasks, as ad hoc and interactive retrieval.

The application to the medical domain is fraught with a variety of challenges which do not had been dealt sufficiently in previous works [5, 6]. Among them, we stand out the following problems. The great part of the summarization systems handles documents wrote in a single language (English, fundamentally), although there are innumerable text collections and resources in other languages (Spanish, specially). Also, most of the systems has been conceived to deal with a restricted subdomain. Therefore, it is necessary to develop techniques that could be applied to broader domains or, at least, that could be easily adaptable from a subdomain to another. As in automatic categorization, we think that the integration of knowledge from resources as UMLS, which has some bilingual components, can play a key role in both problems.

4 The Projects

The SINAMED Project. propose the introduction of original and relevant improvings in the techniques and algorithms, and the specialization and adaptation needed for the specific application environment and the processing of bilingual information (English/Spanish). We are developing an environment for application and experiment of adequate dimension, working with documents of the biomedical domain: Medline, MedlinePlus/HealthDay (English/Spanish) and TREC/Genomics track. This environment integrates text analysis techniques developed with search tools facilitating the information access to the specific user needs. An evaluation of the application environment of each one of the different elements integrated according to general and specific standards of information retrieval, just like the ones used in TREC, and of the concrete operations of text categorization and summarization will be carried out.

The ISIS Project. aims to improve the intelligent access to the medical information, having in mind doctors and patients as end users. It is focused on providing advanced and more effective tools than the current ones for the search, localization, use, and understanding of different sources of medical information. Some interesting aspects are the integrated access to patient's clinical record and related health information.

We intend that, both doctor and patient, exploit the methods and techniques of text mining and intelligent analysis of document's content.

The main scientific and technological objectives of the project are organized around topics as, for example, integration of heterogeneous sources. In this case, the system under development will provide access, in an integrated way, to information coming from the clinical records, scientific articles and others publications concerning health. These sources have very different features as: free text, text endowed with certain external structure (for example, in scientific articles), blended free text with structured data (for example, in clinical results), etc.

The ISIS project has as partners the Universidad Europea de Madrid, Universidad de Huelva, Hospital de Fuenlabrada, a public health care institution with a high technological infrastructure, and Bitex (The bit and text company), a firm specialized in text processing. We are working together in order to decrease the overload of information using text summarization and categorization, improving the organization of answers, presenting groups of related documents and also integrating our algorithms with the SELENE Information System at Hospital de Fuenlabrada.

References

1. Aronson, A.R., Mork, J.G., Gay, C.W., Humphrey, S.M., Rogers, W.J.: The NLM Indexing Initiative's Medical Text Indexer. In: Proceedings of Medinfo, San Francisco (2004)
2. Maña, M.J., de Buenaga, M., Gómez, J.M.: Multidocument summarization: An added value to clustering in interactive retrieval. ACM TOIS, 22 (2), pp. 215-241 (2004)
3. Mostafa J., Lam, W.: Automatic classification using supervised learning in a medical document filtering application. Information Processing and Management 36, 3 (2000) 415-444
4. Ribeiro-Neto B., Laender, A.H.F., De Lima, L.R.: An Experimental Study in Automatically Categorizing Medical Documents. Journal of the American Society for Information Science and Technology 52, 5 (2001) 391-401
5. Elhadad, N., McKeown, K.R..: Towards generating patient specific summaries of medical articles. In: Proceedings of Automatic Summarization Workshop (NAACL), Pittsburgh, USA (2001)
6. Johnson, D.B, Zou, Q., Dionisio, J.D., Liu, V.Z., Chu, W.W.: Modeling medical content for automated summarization. Annals of the New York Academy of Sciences 980 (2002) 47-58

The Universal Object Format – An Archiving and Exchange Format for Digital Objects

Tobias Steinke

German National Library
t.Steinke@d-nb.de

Abstract. Long-term preservation is a complicate and difficult task for a digital library. The key to handle this task is the inclusion of technical metadata. These metadata should be packed together with the files for an exchange between digital archives. Archival systems should handle the data in the Data Management and use it for preservation planning. The German project kopal has defined for this purpose the Universal Object Format (UOF) and enhanced the archival system DIAS with generic functions to support flexible handling of preservation metadata.

Short Description

The German project kopal[1] addresses the problems of the long-term preservation of digital objects. Its goal is the co-operative development of a long-term digital information archive based on the OAIS Reference Model[2]. A key element for such an archive is an open definition of Submission Information Packages (SIP) and Dissemination Information Packages (DIP). A suitable object format should be based on standards, be flexible enough to handle all kinds of digital objects and carry all needed information (metadata) to enable long-term preservation (LTP) strategies. For that, the so-called Universal Object Format[3] (UOF) has been defined. It is based on the Metadata Encoding and Transmission Standard[4] (METS) and Long-term Preservation Metadata for Electronic Resources[5] (LMER). It is flexible enough to include specific technical XML metadata (e. g. the output of the tool JHOVE[6]) and can handle every file format in arbitrary quantities and structures. There is a technical history within the metadata to record migration (file conversion) activities.

Our decision to use METS was based on the fact that METS is very popular in the library community. It is very flexible, but so far it is basically used to store and share the results of digitalisations. We think this schema is also useful for born digital objects and especially the needs of LTP. For this purpose we added special metadata.

[1] http://kopal.langzeitarchivierung.de/index.php.en

[2] http://public.ccsds.org/publications/archive/650x0b1.pdf

[3] http://kopal.langzeitarchivierung.de/downloads/kopal_Universal_Object_Format.pdf

[4] http://www.loc.gov/standards/mets/

[5] http://nbn-resolving.de/?urn=urn:nbn:de:1111-2005051906

[6] http://hul.harvard.edu/jhove/

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 552 – 554, 2006.

At the time of the design of UOF PREMIS[7] had not presented its results. We wanted to have a flexible and practical approach and chose to define our own format based on the data model of the National Library of New Zealand[8]. Even after the release of PREMIS we still think that LMER is a straight forward and practical way to store the needed information for LTP. LMER 1.2 was modularised to fit perfectly in the structure of METS 1.4.

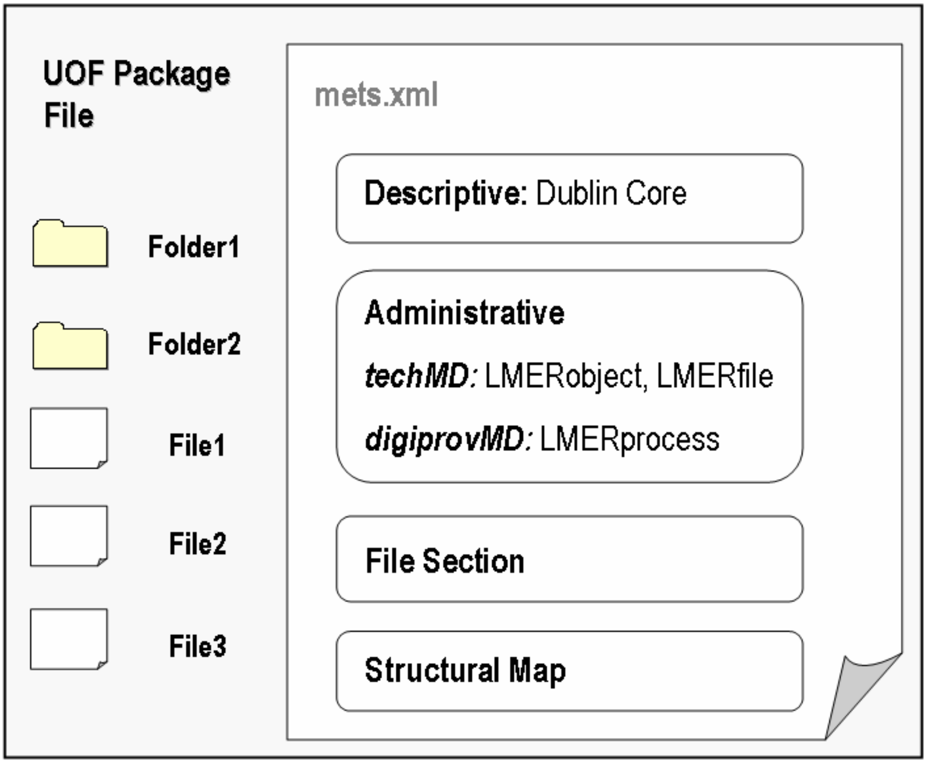


Fig. 1. Generic example of a file in UOF

A package in UOF (figure 1) is one packed file (e.g. with ZIP or tar) containing any quantity of files and any kind of hierarchical file structures. All of these files compose together one logical object, e. g. an electronic theses with three PDF files and one PNG file. There is one file on the root level called mets.xml. This XML file lists in its File Section every file within the package. There is technical metadata for every file and for the complete object. This metadata is stored in the techMD METS section using LMER-File and LMER-Object. If this object is a migration of another object (e. g. the former version was composed of three MS Word files and one GIF file), this history is stored in the digiprovMD METS section using LMER-Process.

[7] http://www.loc.gov/standards/premis/
[8] http://www.natlib.govt.nz/en/whatsnew/4initiatives.html#meta

Depending on the usage of the package there could be also a METS section with descriptive metadata, e. g. in Dublin Core.

The technical information about the digital objects is being stored within the Data Management of the archival system DIAS. DIAS was developed by IBM and enables with its flexible data model the effective use of the data. This data model includes standard data elements and extensible custom data elements to face the challenges of future developments. A query interface to this data is the basis for migration tasks. It should be possible to perform tasks like “Migrate all objects containing b&w-TIFF’s larger than 1 MB, scanned with device XY and ingested before 01/01/2006 to JPEG2000”. Those objects will then be accessed in UOF, migrated and re-ingested in UOF. Both versions of the object will be within in the archive and accessible. We plan to keep at least the first ingested version and the last migration of every object. The project kopal is still going on and we will have a productive archive at the end of 2007.

Tsunami Digital Library

Sayaka Imai[1], Yoshinari Kanamori[1], and Nobuo Shuto[2]

[1] Department of Computer Science, Gunma University,
1-5-1 Tenjin-cho, Kiryu, Gunma 376-8515, Japan
{sayaka, kanamori}@cs.gunma-u.ac.jp
[2] Advanced Research Institute for the Sciences and Humanities,
Nihon University, Ichigaya Tokyu-building 6F, Kudan-Kita 4-2-1,
Chiyoda-ku, Tokyo 102-0073, Japan
shuto-nobuo@arish.nihon-u.ac.jp

Abstract. In this paper, we present our Tsunami Digital Library (TDL) which can store and manage documents about the Tsunami, Tsunami run up simulation, newspaper articles, fieldwork data, etc. We offer a multilingual interface. Currently some documents and explanations of the Tsunami videos have been translated into English and French. We are convinced that TDL will support many people who want to mitigate the Tsunami disaster and to plan countermeasures against the Tsunami.

1 Introduction

In 2004, the world has been struck with the biggest Tsunami in the Indian Ocean. Governments all over the world have since rallied to help developing countermeasures that mitigate such a disaster. In Japan, however, as the archipelago has often been a victim of the Tsunami, there has always been a continuous research on the matter. There are a lot of documents recorded these disasters more than a thousand year. Reports on the Tsunami records date back to the eighth century. In 1960, Japan archipelago has been struck with the big Tsunami from the Chilean Earthquake. Therefore many countermeasures reports have been offered by the Japanese government, many prefecture offices and city offices. Focus on the Tsunami mitigation includes the Tsunami run-up simulation, fieldwork investigation to capture damages in several formats (videos, pictures and descriptive reports), etc. But people who live along the coast, struck with the Tsunami repeatedly, cannot refer these documents easily. In this paper, we present our Tsunami Digital Library (TDL) [1] which can store and manage documents about the Tsunami, the Tsunami run-up simulation, newspaper articles, fieldwork data, etc. And every people who want data about the Tsunami can get information through the internet. We also offer a multilingual interface. Currently some documents and explanations of the Tsunami videos have been translated into English and French. We are convinced that TDL will support many people who want to mitigate the Tsunami disaster and to plan countermeasures against the Tsunami.

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 555 – 558, 2006.

2 Tsunami Digital Library System

2.1 System Overview

We digitalized materials concerning the Tsunami, and made the XML data including Dublin Core. Documents were converted into text by human, for example, handwriting brush records were read by experts, and also type documents were recognized in part by the automatic character recognition system. We stored them into Digital Library. Also we implemented partial documents retrieval from XML documents. Fig. 1 shows the structure of our TDL system. We used Oracle10g Database system to store XML Tsunami data, PostGIS Database system to store fieldwork related data and a video server to store the Tsunami run-up simulation (CG) which simulated and estimated by a past Tsunami records in some regions. We also provided a TDL Portal Server to manage the heterogeneous databases and to allow users to access easily a various types of the Tsunami data.

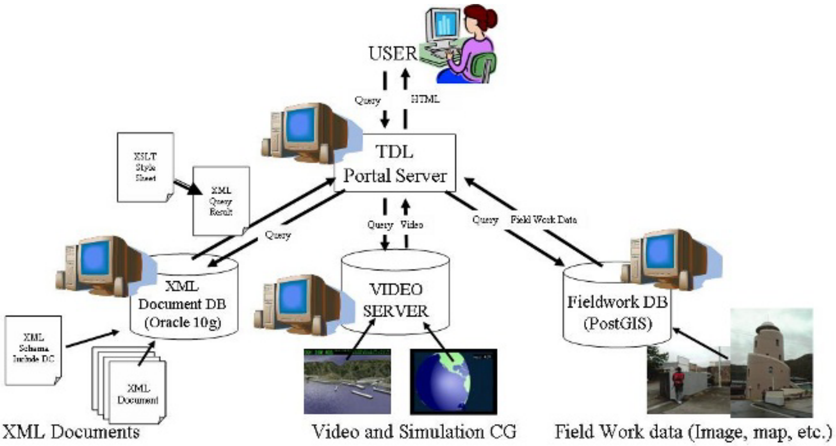


Fig. 1. Outline of the Tsunami Digital Library system

2.2 Contents of Tsunami Digital Library

We collected a various types of data about the Tsunami disaster, such as video, Tsunami run-up simulations (CG) and field work data which were captured by researchers at the Tsunami damage areas. The current TDL contents are follows:

- Reports about the Tsunami by governmental offices or researchers.
  - Four reports for countermeasures against the Tsunami.
  - Eleven reports for the Meiji era earthquake (the Meiji Great Sanriku Tsunami in 1896)

- Twenty-two reports for the Showa era earthquake (the Showa Sanriku Tsunami in 1933, the Showa Tonankai Tsunami in 1944, the Showa Nankai Tsunami in 1946 and the Chilean Tsunami in 1960).

- Six papers about damage of the Tsunami, its mechanism, etc.
- Four experience stories about the Tsunami (the Showa Tonankai Tsunami in 1948).
- Newspapers (articles about the Tsunami).

- Six newspapers published from June 17 to July 14, 1896.

- Eight newspapers published from Mar. 3 to Apr. 30, 1933, from Dec. 7, 1944 to Jan. 31, 1945, and from Dec. 21, 1948 to Feb. 28, 1949.

- Tsunami Run-up Simulations (CG).

- Ten run-up simulations about the Meiji Great Sanriku Tsunami.

- One run-up simulation about the Kanto Earthquake Tsunami in 1923.

- One run-up simulation about the Chilean Tsunami.

- Videos about struck scene by the Tsunami.

- Eight videos about the Chilean Tsunami in 1960, the Tokachioki Earthquake Tsunami in 1968 and the Nihonkai Chubu Earthquake Tsunami in 1983.

2.3 Data Structure of Tsunami XML Documents

In order to show the Tsunami documents effectively, we designed XML Database Schema and XSLT style sheet for the user interface. Fig 2 shows the XML Schema structure of Tsunami XML documents. The Tsunami documents are categorized by types of documents for example newspapers, papers, reports and miscellaneous documents. Fig 2 shows report document of the Tsunami. The root tag is <report> and the document tree consists of <metadata> that represents properties of document and <section> sub tree that represents body of the document. Fig 3 shows an example of XML document data. Fig 4 shows the XML document with XSLT style sheets. As shown in Fig. 4, for example, we can easily read articles according to a table of contents in a document.

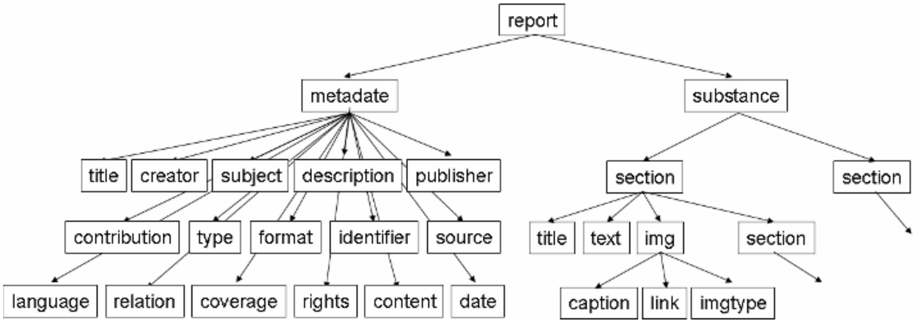


Fig. 2. XML Schema Structure

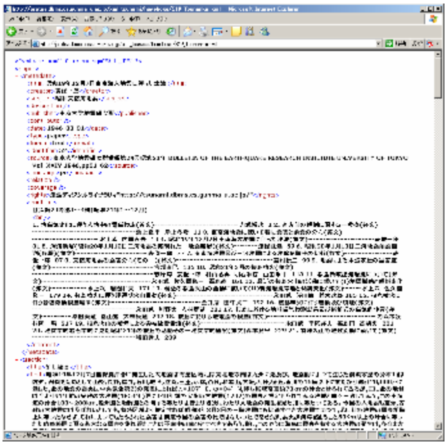
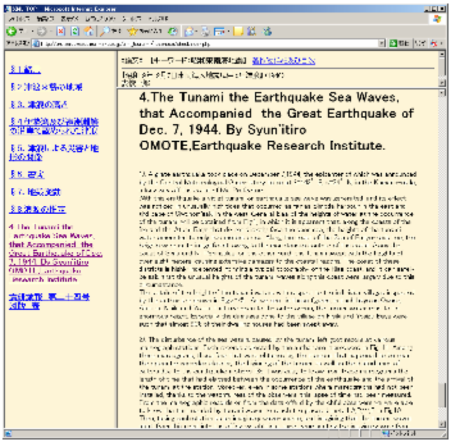


Fig. 3. XML Document



4.The Tunami the Earthquake Sea Waves, that Accompanied the Great Earthquake of Dec. 7, 1944. By Syun'itiro OMOTE,Earthquake Research Institute.

Fig. 4. XML Document with XSLT

3 Conclusion

We have developed the Tsunami Digital Library (TDL) as one of useful applications of the digital library. TDL is constructed by using database systems such as Oracle10g and PostGIS. By using TDL, many people who want information about the Tsunami can get various types of the Tsunami data such as reports, papers, countermeasure documents, newspapers, simulation video and so on. As one of TDL applications we are developing a text book to enlighten the Tsunami disaster based on the contents in the TDL.

Acknowledgments

This work was supported by the Grants-in-Aid for Scientific Research (B), 15300029, 2005, from JSPS and Special Project for Earthquake Disaster Mitigation in Urban Areas, III-3, 2005, and assisted by Ms. Dorsaf Azzabi, PhD student at Gunma University.

References

1. Tsunami Digital Library: http://tsunami.dbms.cs.gunma-u.ac.jp

Weblogs for Higher Education: Implications for Educational Digital Libraries

Yin-Leng Theng and Elaine Lew Yee Wan

School of Communication & Information, Nanyang Technological University
Nanyang Avenue, Singapore 637718
{tyltheng, lewy0001}@ntu.edu.sg

Abstract. Based on a modified Technology Acceptance Model (TAM), the paper describes a study to understand the relationships between perceived usefulness, perceived ease of use and intention to use weblogs for learning in higher education. Data was collected from sixty-eight students of a local university. The findings suggested that students were likely to accept weblog use as a course requirement if they perceived the activity to be useful for learning. The paper concludes with a discussion on design implications for educational digital libraries.

1 Rise of Weblogs in Education

A weblog is essentially a Web page where all writing and editing is managed through a Web browser [1]. The user can publish to the Web without any programming code or server software. Weblog content typically consists of short time-stamped entries arranged in reverse chronological order.

Weblogs offer a number of possibilities for student-centred learning in higher education. They extend the scope for interaction and collaboration among students beyond the physical classroom. Discussion can take place at times and places chosen by students. Interactivity also encourages self and peer assessments, critical aspects in the learning process [2]. Writing entries in weblogs and exchanging ideas with others refine students' thinking and writing skills. Weblogs also support active learning as learning logs track the progress of knowledge construction through all iterations made, rather than simply display finished work [1, etc.]. Weblogs also support the creation of knowledge communities, in which related posts made on disparate weblogs can be connected with hyperlinks [4, etc.].

2 The Study

2.1 Motivation and Theoretical Model

Despite the apparent popularity of weblogs [5], there is a high rate of weblog abandonment after creation. Sifry [5] found that only 13% of all weblogs (currently 1.8 million weblogs) are updated at least weekly. These figures are perhaps not surprising, given the existence of numerous weblog hosting services and the fact that many

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 559 – 562, 2006.

such services enable weblogs to be created quickly, easily and often free of charge by almost anybody with Internet access. However, they also suggest that weblogs may not hold the attention of their users for long. It is thus important to examine students' perceptions of the value of weblogs for learning purposes and the factors that influence those perceptions.

2.2 Study Objectives and Hypotheses

Previous research suggested that TAM might be appropriate for examining students' acceptance of the use of technology for teaching and learning purposes. Using constructs modeled by the well-established Technology Acceptance Model (TAM) [3], this study aimed to investigate the relationships between students' perceptions of weblog usefulness and ease of use and their intentions to use weblogs. The study also examined factors that might influence the acceptance of weblogs as a tool for teaching and learning in higher education, as perceived by university students.

The three main constructs of TAM such as *perceived usefulness* (PU), *perceived ease of use* (PEOU) and *behavioural intention to use* (BI), were incorporated into the theoretical model. PU and PEOU were each proposed as determinants of BI. PEOU was also retained as a determinant of PU. Based on TAM (see Figure 1), a set of hypotheses was generated to answer the research objectives (see Table 1), a subset of which is presented in this paper.

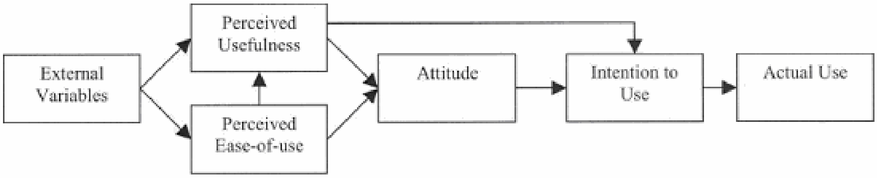


Fig. 1. Technology Acceptance Model

Table 1. Hypotheses 1a, 1b and 1c

Hypothesis 1a:	PEOU has a significant effect on PU.
Hypothesis 1b:	PU has a significant effect on BI to use weblogs as a learning tool.
Hypothesis 1c:	PEOU has a significant effect on BI to use weblogs as a learning tool.

2.3 Questionnaire

A questionnaire instrument was designed to obtain inputs on the eight variables in the model, namely PU, PEOU, BI, and the external variables of PU and PEOU. Respondents were asked to indicate the extent of their agreement with the survey questions using a seven-point Likert-type scale.

Five participants were selected for pre-testing as a means of obtaining feedback on the questions. They were requested to review the questionnaire for ambiguity, repetition, inconsistency, incorrect grammar and any other problems there might be in providing responses to the questions. They were also asked to evaluate the visual appearance of the questionnaire. The questionnaire was revised accordingly.

Since the study was concerned with the perceptions and intentions of university students regarding weblogs, undergraduate and postgraduate students of a local university were selected to participate in the study. A total of sixty-eight students voluntarily participated in the study.

The questionnaire was administered by hand or as an attachment to email messages sent to students on the Master of Information Science degree mailing list maintained by the university. Responses were collected over a period of three weeks from 30 September 2005 to 21 October 2005.

3 Findings and Analysis

3.1 Profiles of Participants and Usage Patterns

94.1% of the respondents were aged between 21 and 40 years with slightly more than half (58%) were male. 88.2 % of the respondents were either Master's or PhD students. Respondents' feedback showed high rates of personal computer ownership (91.2%) and ease of access to personal computers for studies, with 77.9% of the respondents having a computer at home.

It appears that the surveyed population was highly experienced with computing and information technology. An overwhelming majority of respondents rated themselves as highly experienced in operating a personal computer (85.3%), accessing information on the Internet (89.7%), and using email (88.3%). Slightly more than a third (35.3%) of respondents rated themselves highly experienced with weblogs. 23.5% of all respondents reported having their own weblog(s) with 44.8% of all respondents (n=30) counted themselves as weblog readers even though they did not have their own weblogs. Not surprisingly, all who had their own weblog(s) also read weblogs other than their own. Almost a third of the respondents (31.3%, n=21) neither had weblogs nor read them.

3.2 Behavioural Intention to Use (BI): Hypotheses 1a, 1b and 1c

Hypothesis 1a: Effect of PEOU on PU

PEOU had no significant effect on students' PU for learning purposes. The Chi-square value ($\chi^2 = 7.244$) had a significance of 0.124 ($p > 0.05$). In other words, even if students found weblogs easy to use, they might not necessarily consider weblogs a useful learning tool.

Hypothesis 1b: Effect of PU on BI

It was found that PU of weblogs for learning significantly influenced overall intention to use weblogs as a learning tool in higher education ($\chi^2 = 30.839$, $p < 0.001$). Table 2 illustrates the effect of perceptions of weblog usefulness on intentions to use weblogs for a variety of specific learning activities.

Hypothesis 1c: Effect of PEOU on BI

Perceived ease of use of weblogs for learning had no significant influence on overall intention to use weblogs as a learning tool in higher education ($\chi^2 = 1.108$, $p = 0.893$).

No significant results to support Hypothesis 1c in relation to the intention to use weblogs for a variety of specific learning activities, even if the weblogs were perceived to be easy to use.

Table 2. PU and BI for Learning Activities

Effect of PU on BI (Learning Activities)	χ^2	p	**Significant?**
Organize/manage web links	23.400	< 0.001	Yes
Discuss with classmates	21.717	< 0.001	Yes
Self-directed learning	18.418	0.001	Yes
Discuss with tutors	16.644	0.002	Yes
Submit coursework	15.969	0.003	Yes
Work on group projects	14.401	0.006	Yes

4 Implications and Conclusion

A theoretical model was developed based on TAM to better understand the impact of university students' perceptions of the usefulness and ease of use of weblogs on their intentions to use weblogs as a learning tool in higher education. The study found that perceived usefulness was influenced by the awareness of weblog capabilities, peer and tutor support, and students' readiness for interactive learning, which in turn influenced the intention to use weblogs for learning purposes.

The findings suggested that students would likely accept weblog use as a course requirement if they perceived the activity to be useful for learning. If educational DLs were to be truly *dynamic*, allowing user-initiated actions with a social space for collaborative and individual practices, it might be useful for DL designers/developers to learn from the design of weblogs, when creating a dynamic, collaborative, socially-trusted environment in educational DLs.

Acknowledgements

We would like to thank the participants who took part in the survey. This work was carried out as part of NTU's AcRF grant (RG8/03).

References

1. Armstrong, L., Berry, M., & Lamshed, R. (2004). Blogs as Electronic Learning Journals. *e-Journal of Instructional Science and Technology, 7*(1).
2. Connell, S. (2004). *Uses for Social Software in Education: A Literature Review*. Retrieved August 20th, 2005, from http://soozzone.com/690review.htm
3. Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use and User Acceptance of Information Technology. *MIS Quarterly, 13*, 319-340.
4. Oravec, J. A. (2003). Blended by Blogging: weblogs in blended learning initiatives. *Journal of Educational Media, 28*(2-3), 225-233.
5. Sifry, D. (2005, August 2rd). *State of the Blogosphere, August 2005, Part 1: Blog Growth.* Retrieved August 4th, 2005, from The Technorati Weblog at http://www.technorati.com/weblog/2005/08/34.html

XWebMapper: A Web-Based Tool for Transforming XML Documents

Manel Llavador and José H. Canós

Dep. of Computer Science (DSIC)
Technical University of Valencia
46022 – Valencia, Spain
{mllavador, jhcanos}@dsic.upv.es

1 Introduction

Interoperability has been one of the most challenging issues of last decade. Different solutions with various levels of sophistication have been proposed, such as wrappers, mediators, and other types of middleware. In most solutions, the Extensible Markup Language (XML) has been accepted as the de facto standard for the interchange of information due to its simplicity and flexibility.

XML has been particularly successful in the field of Digital Libraries (DLs), where most interoperability problems come from the heterogeneity of metadata formats. By using XML as the language to exchange records, and its associated transformation language XSLT, it is very easy to convert records from one format to another. The only requirement is to have a well-defined transformation in terms of a XSLT template, the construction of which is not always easy.

We faced this problem during the development of Bibshare, an environment for bibliography management that allows users to collect bibliographic references, insert citations into documents, and automatically generate a document's bibliography. Unlike previous tools, Bibshare works with a variety of word-processing systems and permits references to be inserted not only from personal citation collections but also from bibliography servers available on the Internet, including the Open Archives Initiative (OAI) data providers. As expected, each collection has its own metadata format(s). In order to unify the resulting sets of federated searches and return these data to the user, each retrieved record must be converted to a common format, which we call the Bibshare Bibliographic Format (BBF). Assuming all records are encoded in XML, the XSLT template for converting records in a given collection to the BBF must be provided by the manager of the collection. However, many times the expertise of the manager is too limited to deal with XSL, so there is a clear need for tools that support the conversion process.

There are several tools that allow XML document transformation via the generation of XSL templates. One group includes the converting tools that are part of larger suites, especially frameworks for the specification and execution of business processes in e-commerce applications; for example, Oracle BPEL [1] and Microsoft BizTalk Server [2] include XSL editors. The main drawback of these tools is that, in order to have transformation capabilities, the whole suite must be installed, which is sometimes very expensive. Another group includes the set of stand-alone converting tools, such as Altova MapForce [3]. Tools like this define transformations between XML schemas using

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 563 – 566, 2006.

visual metaphors. However, the user must know the source and target schemas in depth, because otherwise some elements cannot be included in the mappings. Stylus Studio [4] allows the definition of very complex mappings, but the price to pay is a complicated user interface that requires in-depth knowledge of XSL.

Some specific-purpose tools allow the rapid generation of transformation templates from/to specific XML schemas (e.g. from XML to HTML). These include Microsoft InfoPath [5], or the aforementioned Stylus Studio, that produces XML documents from databases, Web services, semi structured files, and vice-versa. However, target templates are sometimes hard-coded and transformation rules are precompiled. There are also sophisticated frameworks that allow the automatic definition of the mappings, like XMapper [6], or that at least suggest possible mappings that must be confirmed by the user, as is the case with Schema Mapper [7]. These tools use different formalisms, ontologies, or simply syntactic proximity relationships to map elements of a source schema into elements of a target schema. These approaches are efficient for simple cases; however, most of them have limitations regarding the structure of the documents to transform or the type of relationships between elements; moreover, they do not allow the definition of data transformation functions.

In this work, we introduce XWebMapper, a tool that allows the (semi)automatic generation of XSLT templates. Although it was developed for the Bibshare project, it is a general purpose tool as it corresponds to a generic language such as XML. Given two XML schemas S1 and S2, XWebMapper obtains the XSLT template transforming S1-valid documents into S2-valid documents using a set of semantic mappings that, at the current state of the tool, must be defined by the user in a very intuitive way. Since XWebMapper was implemented as a Web application, its software and hardware requirements are lighter. XWebMapper hides the complexity of source and target schemas using the idea of “concept” to group all the elements with the same name and obtaining the corresponding XPath expressions automatically. XWebMapper also provides a visual interface to define complex mappings. Its implementation based on service-oriented paradigm allows its different components to be called from different applications, as is the case for Bibshare.

2 Workflow and Arhitecture

In this section, we explain the process of transforming documents using XWebMapper, as well as its architecture. XWebmapper can be used from the URL http://bibshare.dsic.upv.es/XWebMapper. Clicking on the “Execute XWebMapper” button starts the generation process (summarized in Figure 1 (a)). The different steps of the process and the components of the arquitecture that support each step (Figure 1 (b)) are explained below.

- **XPath Expressions Inference:** The first step of the process consists of identifying the elements and attributes of the source and target XML schemas, as well as their locations in the form of XPath expressions. This is done automatically by means of the XPathInferer Web Service[1] which makes a recursive exploration of the structure of the schemas. When a schema is not

[1] http://bibshare.dsic.upv.es/XPathInferer/XPIWS.asmx

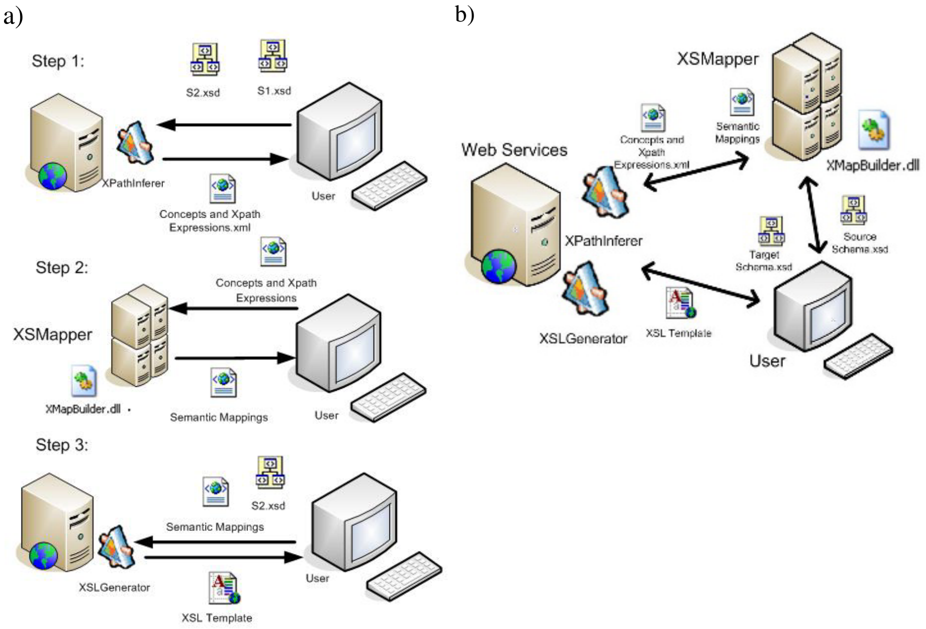


Fig. 1. (a) XSL template definition phases and (b) Service-Oriented Architecture of the Framework

available, a sample document can be used to automatically infer a schema which the document is compliant with.

- **Mapping definition:** After the elements and attributes detection, the process of defining the relationships starts. There are two types of semantic relationships, namely direct relationships and data transforming relationships. The former allow copying the value of an element or attribute of the source schema to an element or attribute of the target schema without any change, whereas the latter allow the modification of values by means of transformation functions. Both types of relationships are defined using the XWebMapper toolbar. Being our aim to be fully compliant with the XSL specification, we have used the functions included in the XSL and XPath languages, which include string, arithmetic, logic, and navigation functions. The XMapBuilder.dll library includes two types of components: Windows Controls, and Classes for XML persistence. The Windows Control Library includes the user's interface of the application. It is implemented with the Microsoft .NET Framework, that allows both the use of controls of this type in Web environments through the conventional web browsers and the downloading from the server and execution in the user’s computer (do not require installation of any type, and they are early updated as the updating is done only in the server and not in the client). The XML persistence classes store the relationships between source and target schemas in XML format[2].

[2] http://bibshare.dsic.upv.es/relations.xsd

- **XSL Generation:** After the definition of the relationships, the final step consists on the automatic generation of the XSL template . This step is supported by XSLGenerator[3]. This Web Service takes as input the schema of the target document, wich is used to construct the syntactic structures of the XSL template, and the set of semantic relationships, that are used to construct the data-selection structures to inject data of the source document into the result.

3 Conclusions and Further Work

Converting XML documents is a very common need in most distributed applications. XSLT is the technology created to automate the conversion tasks, but there the difficulty of generating the transformation templates remains. In this paper, we have introduced XWebMapper, a Web-based environment that uses semantic relationships between concepts to automate the creation of XSLT templates. Although its development was motivated by a specific need in the Bibshare project, it can be considered to be a general solution to the problem of transformation of XML documents.

Unlike previous solutions, the Web interface of XWebMapper has very few requirements. Moreover, users do not have to worry about the possible updates of its components. Its service-oriented architecture allows its components to be invoked by third party applications. Finally, since our approach is based on elements and attributes rather than on the structure of the documents, the user interaction is simplified and there is no need to know the XPath language in depth.

We are currently improving the user interface components. As further work, we plan to add new capabilities for the automatic generation of the transformation templates using ontologies. We also want to exploit its web-based nature to create a catalogue of transformation templates with the help of all its users.

References

1. Oracle BPEL Process Manager.www.oracle.com/technology/products/ias/bpel/index.html
2. Microsoft BizTalk Server: Home. www.microsoft.com/biztalk/
3. Altova MapForce. www.Altova.com/MapForce
4. Stylus Studio. www.stylusstudio.com/xml_product_index.html
5. Microsoft Office Online: InfoPath 2003 Home Page. http://office.microsoft.com/infopath/
6. Kurgan, L., Swiercz, W., and Cios, K. *Semantic Mapping of XML Tags using Inductive Machine Learning*. http://citesser.ifi.unizh.ch/kurgan02semantic.html
7. Raghavan, A., Rangarajan, D., Shen, R., Gonçalves, M.A., Srinivas, N., Fan, W., and Fox, E. *Schema Mapper: A Visualization Tool for Digital Library Integration*. Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital libraries, ACM

[3] http://bibshare.dsic.upv.es/XSLGenerator/XSLGeneratorWS.asmx

Author Index