DELOS
Network of Excellence on Digital Libraries

DELOS Research Activities 2006

Editor: Costantino Thanos (ISTI-CNR)

August 2006

DELOS: a Network of Excellence on Digital Libraries
www.delos.info

Information Society Technologies

Sixth Framework Programme

DELOS

Network of Excellence on

Digital Libraries

Information Society Technologies

Table of contents

Introduction

DELOS is a Network of Excellence on Digital Libraries partially funded by the European Commission in the frame of the Information Society Technologies Programme. It started on 1st January 2004 and has a duration of 48 months. It presently has 56 members.

Digital Libraries represent a new infrastructure and environment that has been made possible by the integration and use of a number of IC technologies, the availability of digital content on a global scale and by a strong demand from users who are now on-line. They are destined to become an essential part of the information infrastructure in the 21st century. The DELOS vision for digital libraries is that they should enable any citizen to access all human knowledge any time and anywhere, in a friendly, multi-modal, efficient and effective way, by overcoming barriers of distance, language, and culture and by using multiple Internet-connected devices. The new generation digital libraries should not just be seen as static information repositories but as growing, interactively, and collaboratively used nuclei of what will be at some stage, a good part of human knowledge that depends as much on information as on communication.

Digital Libraries are starting to support the specialized needs of very diverse technologies and applications, from cultural heritage to general science, health, government, and education. After approximately ten years of development, they have moved far beyond any connotations of the term "Library", to also encompass Digital Archives and Museums and now have functionality to deal with multimedia objects often with embedded general knowledge, semantics, and behaviour. To fulfil their new roles as universal knowledge infrastructures, Digital Libraries require research in several new key areas pointing to the development of:

- user-centered system design methodology
- pro-active systems with functionality that facilitates collaboration, communication, and information creation
- generic Digital Library Management Systems that provide basic system infrastructures that can be used to implement application specific digital libraries incorporating context-specific services.

The DELOS Network of Excellence in Digital Libraries intends to advance the field in these new and exciting directions, with the aim of progressing to the development of the next-generation Digital Library system. To this end, DELOS coordinates a joint programme of activities of the major European teams working in digital library related areas. The objective is to develop dynamic universal knowledge environments, which will transform research and education at all levels by collecting, organizing and making publicly accessible on-line vast quantities of information. The ultimate goal is to provide access to human knowledge from anywhere and any time and in an efficient and user-friendly fashion. DELOS also aims at disseminating knowledge of digital library technologies to many diverse application domains, by providing access to technological know-how, services, test-beds, and the necessary expertise to facilitate their take-up. In this context, one of the major efforts started last year is the cooperation with the Office of The European Library, aimed at transferring digital library functionality developed by DELOS members to the TEL system.

The research activities of DELOS have been organized in seven clusters:

- Digital Library Architecture
- Information Access and Personalization
- Audio/Visual and Non-traditional Objects
- User Interfaces and Visualization
- Knowledge Extraction and Semantic Interoperability
- Preservation
- Evaluation

In addition, there is an eighth cluster gathering all the activities related to visibility, dissemination, education and technology transfer or, to use the EU terminology, dedicated to the “Spreading of Excellence”.

In the following pages the major projects being carried on in each cluster are briefly described. They have also been shown in a poster session held at ECDL 2006 in Alicante. For additional information about DELOS please visit www.delos.info. For further information about specific research projects, please get in contact with the project coordinators.

Digital Library Architecture

Cluster objectives

Citizens of the future should be able, through the medium of better designed digital libraries to gain access to a myriad of forms of knowledge from anywhere and at any time and in an efficient and user-friendly fashion. But for this to happen those digital libraries will need to arrive at a common infrastructure which is highly scalable, customizable and adaptive. From a technical viewpoint, this infrastructure has to support state-of-the-art and promising innovative models and techniques, and frameworks to develop and evaluate digital libraries, and has to be highly customizable, configurable and adaptive. To this end, various activities and developments have to be seamlessly integrated into a coherent whole to develop such a generic and modular digital library infrastructure. This includes architectural approaches, processes, and activities.

Architecture. Peer-to-peer architectures allow for loosely coupled integration of information services and sharing of information such as recommendations and annotations. Different aspects of peer-to-peer systems (e.g. indexes, and P2P application platforms) must be combined. Grid computing middleware is needed because certain services within digital libraries are complex and computationally intensive (e.g., extraction of features in multimedia documents to support content-based similarity search or for information mining in bio-medical data). The service-oriented architecture provides mechanisms to describe the semantics and usage of information services. Moreover, it supports mechanisms to combine services into workflow processes for sophisticated search and maintenance of dependencies.

Processes. Applications within digital libraries must consider the autonomy and distribution of information providers. Hence, accessing information means combining existing services into composite applications, i.e. workflow processes. The main topics to be considered are therefore Workflow Management, Publish/Subscribe Techniques, Replication and Freshness of Data, Mobile Information Components.

Functions. Data and documents within a digital library are made available by dedicated services. These services allow for the definition of building blocks, which are tailored to the type of data and documents and implement, for instance, appropriate index structures. The main types of data to be considered are XML data Storage and Access, Multimedia data Access, Digital Rights Management, Security and Certification.

The overall goal of this workpackage is to analyze, develop, and integrate architectures and technology for digital libraries that enable the building of the next generation digital library management systems. The ambitious objective is to have a demonstrator for future Digital Library Management Systems that not only shows new combined text and audio-visual search functionality and personalized browsing by new adaptable information visualization and relevance feedback tools at the interface but also proves that generic systems can be build that not only support finding of relevant information but also enable to annotate and process found information, to integrate sensor data stream processing, and – from a systems engineering point of view - allows simple configuration and adaptation while being reliable and scalable.

Cluster activities

To achieve the goals sketched above, the cluster will concentrate on the following activities:

- Organisation of workshops on digital library architectures
- Developments of surveys that collect most significant contributions of peer-to-peer data management, grid computing, and service-orientation for digital library architectures
- Evaluation of approaches to connection management and information synchronisation
- Developments of prototype software modules and components for web services, multiple service composition and management, wireless connectivity
- Development of a benchmark for the evaluation of digital library architectures
- Test of the solutions on a prototype ongoing application.

Cluster coordinator

Hans-Jörg Schek, University of Konstanz, Germany, schek@inf.ethz.ch

A Reference Model for Digital Libraries

Donatella Castelli, ISTI-CNR, donatella.castelli@isti.cnr.it
Yannis Ioannidis, National and Kapodistrian University of Athens, yannis@di.uoa.gr
Seamus Ross, HATII, University of Glagow, s.ross@hatii.arts.gla.ac.uk
Hans Joerg Schek, University of Konstanz, schek@inf.ethz.ch
Heiko Schuldt, University of Basel, heiko.schuldt@unibas.ch

Keywords: H.3.7: Digital Libraries, H.1.1: System and Information Theory

Objectives

Many attempts were made in the past to define what a digital library (DL) is [1,3,6]. However, ,over the time this notion has evolved a lot, mainly because of the stimulating contributions by the DELOS community[4,5]. Today, it is has become obvious that “Digital Library” is a complex notion with several diverse aspects that cannot be captured by a simple definition. A comprehensive representation encapsulating all potential perspectives is therefore required. The objective of the DELOS Task 1.4 is to satisfy this demand by systematically introducing appropriate models of the DL Universe. The ultimate goal of this activity is to lay the foundation of the DL field and to stimulate the definition of the related systems standards.

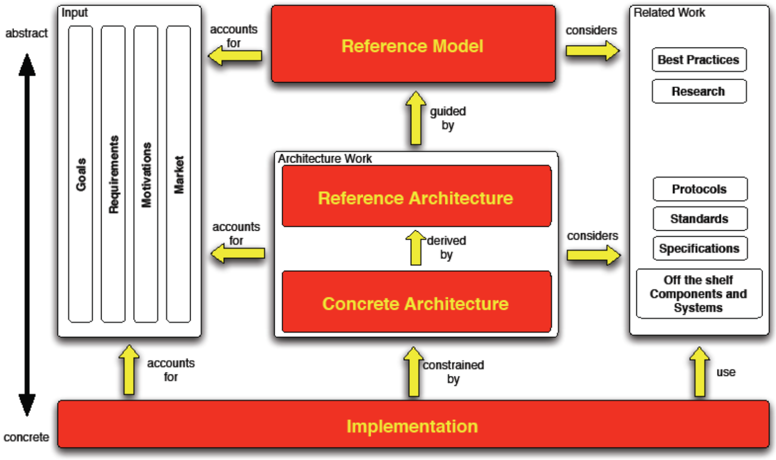


Figure 1- The Digital Library framework

Outcomes

Figure 1[1] illustrates the DL reference framework and its relations with the outside world as identified within Task 1.4. The activity performed in this task has been focussed in particular on two of the four main components highlighted in red: the Reference Model and the Reference Architecture. Three documents, now available on the project website (http://www.delos.info/index.php?option=com_content&task=view&id=345), present the outcomes of this activity:

The *Digital Library Manifesto* - introduces the DELOS view of the DL universe and of its modelling needs. This view has been obtained by leveraging on the results of several European research groups active in the Digital Library field for many years, both within the DELOS Network of Excellence and outside, as well as on results of other groups around the world. It is to serve as a springboard for future work. In particular, the Manifesto introduces the necessary distinct notions of “system” in the DL Universe, namely, *Digital Library Management System*, *Digital Library System*, and *Digital Library*. Although these systems are often confused and used interchangeably, each of them plays a central and distinct role in the DL universe and has, therefore, been explicitly defined. The Manifesto also identifies six main concepts that characterise dimensions according to which the modelling of the systems can be organised, namely, *Information Domain*, *Functionality*, *User*,

[1] This picture has been inspired by a similar one reported in [2]

Quality, *Policy* and *Architecture* (see Figure 2). Finally, the Manifesto highlights four main different roles that actors may play with respect to a Digital Library, *DL End-Users*, *DL Designers*, *DL System Administrators*, and *DL Application Developers*. Each role is primarily associated with one of the three "systems" (layers of abstraction) and, therefore, has different modelling needs.

The *DL Reference Model* – introduces a minimal set of unifying inter-related concepts that collectively circumscribe and capture the essence of the DL notions. It is independent of specific standards, technologies, implementations, or other concrete details. It considers the three systems identified in the Manifesto and models them according the six identified dimensions.

A *DL Reference Architecture* – presents an architectural design pattern indicating an abstract solution for implementing the concepts and relationships identified in the Reference Model. There may be more than one reference architecture that addresses how to design a DL, the difference laying in the requirements imposed by the application scenario. The proposed reference architecture assumes a scenario where (i) the organizational model is a federated one, based on resources sharing, (ii) the number of participating organizations can (dynamically) grow and (iii) the user requirements can evolve over the time. As a consequence of these requirements the proposed reference architecture is component-based.

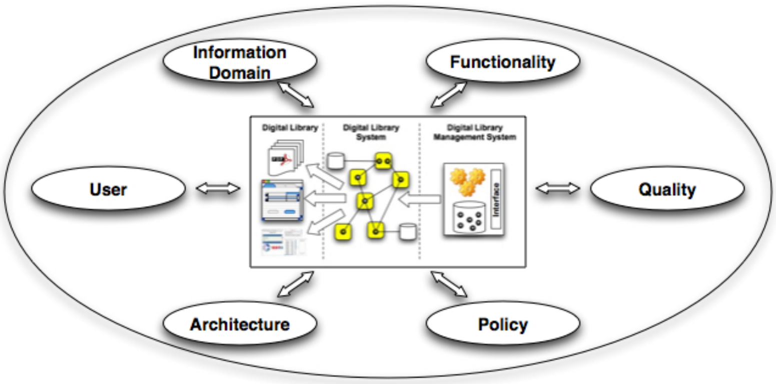


Figure 2 – The digital library main concepts

Plans for the future

The three documents have been presented in an International Workshop that was held in Frascati on June 2006. The feedback received from the participants has stimulated enrichments and revisions. The new version of these documents will be distributed again to the workshop participants, some of which are willing to further contribute to the refinement of the presented models.

References

[1] Borgman C. L. 1999. What are digital libraries? Competing visions. In *Information, Processing and Management* 35(3): 227-243.

[2] Duane N. 2005. *Service Oriented Architecture*. White paper. Adobe Systems Inc.

[3] Fox E. A. and Marchionini G. 1998. Toward a worldwide digital library. In *Communications of the ACM* 41(4): 29–32.

[4] Ioannidis Y. (ed.) 2001. Digital Libraries: Future Directions for a European Research Programme. DELOS Brainstorming Report.

[5] Ioannidis Y. 2005. Digital libraries at a crossroads. In *International Journal of Digital Libraries* 5(4): 255-265.

[6] Soergel D. 2002. A framework for digital library research: broadening the vision. In *D-Lib Magazine* 8(12).

Management of and Access to Virtual Electronic Health Records2

Peter M. Fischer, Swiss Federal Institute of Technology (ETH) Zürich, peter.fischer@inf.ethz.ch
Yannis Ioannidis, National and Kapodistrian University of Athens, yannis@di.uoa.gr
Hans Joerg Schek, University of Konstanz, schek@inf.ethz.ch
Heiko Schuldt, University of Basel, heiko.schuldt@unibas.ch
Michael Springmann, University of Basel, Michael.springmann@unibas.ch
Ulrike Steffens, Kuratorium OFFIS, ulrike.steffens @offis.de
Raimund Vogl, Health Information Technologies Tyrol (HITT), raimund.vogl@hitt.at

Introduction

eHealth digital libraries contain electronic artifacts that are generated by different healthcare providers (family doctors, laboratories, hospitals, etc.). An important observation is that this information is not stored at one central instance but rather under the control of the organization where data has been produced. The electronic health record of patients therefore consists of a set of distributed artifacts and cannot be materialized for organizational reasons. Rather, the electronic patient record is a virtual entity and has to be generated by composing the required artifacts each time it is accessed. The virtual integration of an electronic patient record is done by encompassing services provided by specialized application systems into processes. A process to access a virtual electronic health record encompasses all the services needed to locate the different artifacts, to make data from the different healthcare providers available –given appropriate authorization and authentication–, to perform the format conversions needed, and to present the (possibly anonymized) result to a user (i.e., patient X accesses his virtual health record via web; it contains collected medical documents from different health care institutions. Or doctor W, a specialist for internal medicine, retrieves the automatically integrated information related to the coronary heart disease of his patient Y from several health care institutions). In addition to the patient-centric access to virtual health records, also disease-specific applications can be supported by means of processes. These disease-specific applications allow for epidemiological studies across a set of patients, comparisons, identification of similar diagnoses, etc. (i.e., medical scientist M would like to identify all patients that have similar pathological deviations in the X-ray of their lung than patient Z, for whom SARS has been diagnosed. Or the Ministry of Health needs a statistical overview about diagnoses and treatments during the last three years, in order to control the health system and –for civil protection– to prevent from dangerous sanitary situations).

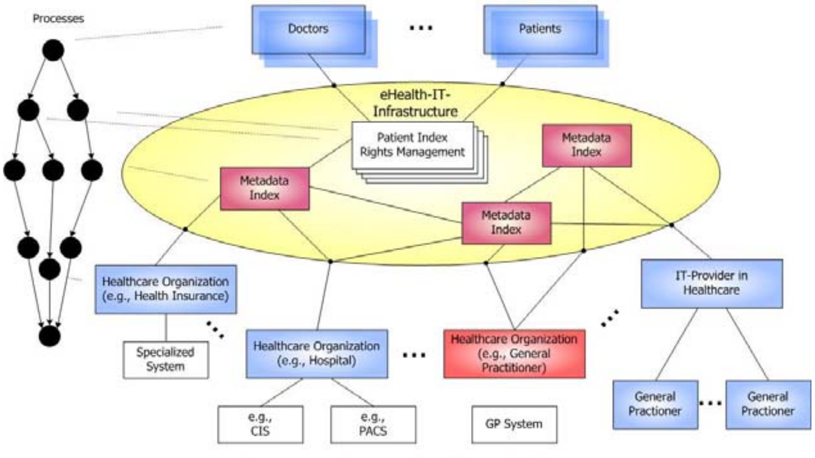

Processes
Doctors
...
Patients
eHealth-IT-Infrastructure
Patient Index
Rights Management
Metadata Index
Metadata Index
Metadata Index
Healthcare Organization (e.g., Health Insurance)
Specialized System
Healthcare Organization (e.g., Hospital)
...
Healthcare Organization (e.g., General Practitioner)
IT-Provider in Healthcare
e.g., CIS
e.g., PACS
GP System
General Practioner
...
General Practioner

Figure 1: System Architecture for Virtual Electronic Health Records

Research Objectives

The realization of these goals first requires the availability of appropriate (web) services in order to access relevant data managed by specialized applications which are hosted by different healthcare organizations. In addition, common standards have to be supported to integrate these legacy applications (e.g., the PACS application where the X-rays of Y and Z are stored) or, alternatively, dedicated services are needed to transform the format of the data retrieved from one application so as to make is available for other, subsequent services. Second, an infrastructure to combine these services into processes is needed that is highly dependable and reliable. Physicians must be given the guarantee that the system and data is always available (i.e., by means of replication) and that processes come to a well-defined end (e.g., by collecting all pieces of information that are of interest), even in case of failures.

[2] This joint research is supported by the EU in the 6th Framework Programme within the Network of Excellence DELOS (contract no. G038-507618).

Third, the infrastructure has to provide a high degree of scalability, and to efficiently schedule the access to computationally intensive services by applying sophisticated load balancing strategies using grid technology. Physicians need information immediately, especially for patient-centric queries (but also for certain disease-specific queries), in order to make vital decisions. Hence, long response times due to a high system load cannot be tolerated. Consider the first disease-specific scenario above (the lung X-ray of patient Z) where similarity search across a potentially large set of documents is needed. In order to support this search, feature extraction has to take place for all documents/images, nearest neighbours have to be determined, etc. All these steps require significant computing power and should not be limited to the organization where the images are stored. Rather, additional feature extraction services should be installed automatically at hosts which currently feature a low load. Finally, the infrastructure has to allow for the transparent access to distributed data by means of appropriate (peer-to-peer) indexing techniques that avoid single points of failures as well as censorship and that, at the same time, preserve the privacy of data. The systems architecture of an infrastructure to support virtual electronic health records is depicted in Figure 1.

Expected Results

The main goal of this DELOS task is to identify, design, and build demonstrators for the basic building blocks needed to access virtual electronic health records, i.e., locate the different artifacts, make data from the different healthcare providers available, perform the format conversations needed, and present the result to a user. The demonstrator system which has been implemented and which is currently being extended shows extensions to the OSIRIS/ISIS prototype system (UNIBAS) for content-based search in medical image collections, the XL system (ETH) for the definition and deployment of new web services (which provide functionality to be used in a eHealth setting like integration of results combing from different healthcare providers or displaying query results as web pages) and eaSim and ORCA (OFFIS). eaSim is a simulator for peer-to-peer environments. In the context of an eHealth DL, this is particularly important to simulate query processing in a distributed environment as it is shown in Figure 1. ORCA supports the modelling of organizational structures and also the definition of roles and access rights to fragments of an electronic health record. Finally, a first version of an infrastructure for a regional, shared electronic patient record in the Austrian state Tyrol has been implemented (HITT).

Project Organization

The authors of this extended abstract are the members of this project. The partner institutions are:
Health Information Technologies Tyrol (HITT), Innsbruck, Austria
Kuratorium OFFIS, Oldenburg, Germany
National and Kapodistrian University of Athens, Greece
Swiss Federal Institute of Technology (ETH) Zürich, Switzerland
University of Konstanz, Germany
University of Basel, Switzerland

References

[1] Bischofs L., Hasselbring W., Niemann H., Schuldt H. and Wurz M. 2004. Verteilte architekturen zur intra- und inter-institutionellen integration von patientendaten. In E. Ammenwerth, W. Gaus, R. Haux, C. Lovis, K.P. Pfeiffer, B. Tilg and H.E. Wichmann (eds.), *Kooperative Versorgung, Vernetzte Forschung, Ubiquitäre Information. Tagungsband der 49. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS 2004). Innsbruck (Austria), 26-30 September 2004*. Niebüll: Videel Verlag OHG. 87-89.

[2] Florescu D., Grünhagen A., Kossmann D. 2003. XL: a platform for web services. In CIDR 2003, First Biennial Conference on Innovative Data Systems Research. Asilomar (USA), 5-8 January 2003.

[3] Grimson J., Grimson W. and Hasselbring W. 2000. The SI challenge in health care. In *Communications of the ACM* 43(6): 48-55.

[4] Hasselbring W. 1997. Federated integration of replicated information within hospitals. In *International Journal on Digital Libraries* 3(1): 192-208.

[5] Pedersen S. and Hasselbring W. 2004. Interoperabilität für informationssysteme im gesundheitswesen. In *Informatik Forschung und Entwicklung* 18(3): 174-188.

[6] Schabetsberger T., Ammenwerth E., Goebel G., Lechleitner G., Penz R., Vogl R. and Wozak F. 2005. What are functional requirements of future shared electronic health records? In R. Engelbrecht, A. Geissbuhler, C. Lovis and G. Mihalas (eds.). *European Notes in Medical Informatics (CD-Rom): Connecting Medical Informatics and Bio-Informatics (MIE 2005). Geneve (Switzerland), 28-31 August 2005*. Vol. 1(1).

[7] Springmann M. 2006. A novel approach for compound document matching. In *Bulletin of the IEEE Technical Committee on Digital Libraries* 2(2).

[8] Springmann M., Schek H.-J. and Schuldt H. 2004. Kombination von Bausteinen zur ähnlichkeitsbasierten Suche in elektronischen Multimedia-Patientenakten. In E. Ammenwerth, W. Gaus, R. Haux, C. Lovis, K.P. Pfeiffer, B. Tilg and H.E. Wichmann (eds.) *Kooperative Versorgung, Vernetzte Forschung, Ubiquitäre Information. Tagungsband der 49. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS 2004). Innsbruck (Austria), 26-30 September 2004*. Niebüll: Videel Verlag OHG.

Integration of Data Stream Management into a Digital Library[3]

Gert Brettlecker, University of Basel, Gert.Brettlecker@unibas.ch
Peter M. Fischer, Swiss Federal Institute of Technology (ETH) Zürich, peter.fischer@inf.ethz.ch
Yannis Ioannidis, National and Kapodistrian University of Athens, yannis@di.uoa.gr
Hans Joerg Schek, University of Konstanz, schek@inf.ethz.ch
Heiko Schuldt, University of Basel, heiko.schuldt@unibas.ch

Introduction

Recent trends in ubiquitous and pervasive computing, together with new sensor technologies, wireless communication standards, powerful mobile devices and wearable computers strongly support novel types of applications. An example is tele-monitoring in healthcare which makes use of this new technology in order to improve the quality of treatment and care for patients and the elderly. In particular, if we consider our aging society, the amount of elderly people suffering from one or more chronic diseases is increasing. Tele-monitoring applications enable healthcare institutions to take care of their patients while they are out of hospital, which is especially useful for managing various chronic diseases as well as for measuring the effects of treatments under real-life conditions. A similar but even more comprehensive application is the support for the elderly and for people with cognitive disabilities living at home. Here, not only physiological sensors and data are relevant for tele-monitoring but also context information, e.g., information on the current activities of a person, where she or he is located, etc. In the DELOS task Integration of Data Stream Management into a Digital Library, different stream applications will be considered, yet with particular focus on eHealth.

Research Objectives

Continuous data streams generated by (wearable) sensors have to be processed online in order to detect critical situations. This is important for instance for traffic or network monitoring, but also for eHealth applications. For this purpose, usually different streams (generated by different types of sensors) have to be combined (e.g., jointly consider oxygen saturation, ECG signals and blood pressure). This is done by making use of specialized operators (e.g., signal pre-processing, filtering, joining different streams, aggregating stream data, and feedback looping etc.). In Figure 1, a concrete setting taken from tele-monitoring in eHealth is depicted. Such an infrastructure has to be able to combine these operators in an application-specific way. However, in addition to the stream operators, also traditional discrete (web) services, e.g., services that do not operate on continuous input data, have to be integrated. These services are needed, for instance, in order to notify emergency physicians in very critical cases, to notify neighbors about abnormal situations, but also to write information aggregated from stream data back to the electronic health record of a patient. The latter is very important to integrate data stream management into an eHealth digital library. Reliability of the eHealth digital library infrastructure is of utmost importance. In order to achieve a high degree of reliability, the task addresses failure handling a various levels from message level,

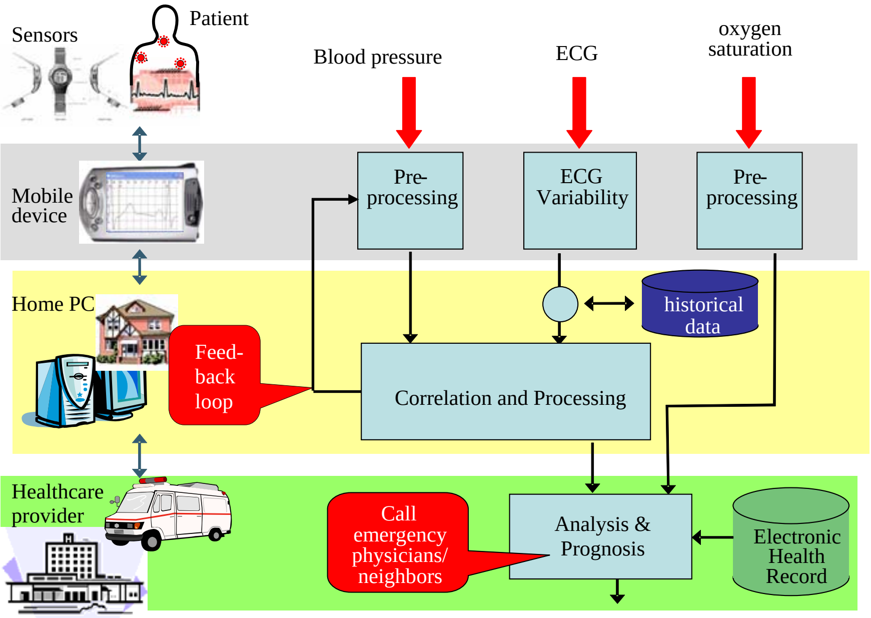


Figure 1: A Sample Data Stream Process for Health Monitoring

[3] This joint research is supported by the EU in the 6th Framework Programme within the Network of Excellence DELOS (contract no. G038-507618).

operator or activity level, to process level. In particular, at operator level the task provides sophisticated and efficient operator checkpointing in order to allow for application-transparent operator migration.

Expected Results

The main goal of this task is to identify, design, and build demonstrators for data stream operators, i.e., the continuous processing of combined streaming data generated by different types of sensors. Appropriate web services are also needed in order to process and store the results and aggregates of stream operators. Finally, appropriate notification mechanisms are needed in order to inform a patient as well as healthcare providers on critical deviations of her/his health state. These operators and services will be combined in an infrastructure for stream processing. While in the first year of this task, the focus has been exclusively on eHealth applications, its scope has been generalized. Currently, also other application domains (e.g., traffic management) are being addressed. In particular, this task started with an evaluation of applications where the combination of data stream operators (continuous processing of streaming data) and web services (discrete invocation of application functionality) is an inherent requirement. Based on this evaluation, selected streaming operators (e.g., for searching outliers in a data stream or for joining parallel data streams) are currently being implemented and notification mechanisms will be provided. In addition, a prototype implementation of an infrastructure for workflow processes including stream processing supporting the integration of stream operators and web services will be provided. This allows for the insertion of the results of data stream processing into a DL, thereby tightly integrating data stream processing and Digital Library Systems.

Project Organization

The authors of this extended abstract are the members of this project. The partner institutions are:

National and Kapodistrian University of Athens, Greece; Swiss Federal Institute of Technology (ETH) Zürich, Switzerland; University of Konstanz, Germany; University of Basel, Switzerland

References

[1] Brettlecker G., Schuldt H. and Schek H.-J. forthcoming. Efficient and coordinated checkpointing for reliable distributed data stream management. In *Proceedings of the 10th East European Conference on Advances in Database and Information Systems (ADBIS 2006). Thessaloniki (Greece), 3-7 September 2006*.

[2] Brettlecker G., Schuldt H. and Schek H.-J. 2005. Towards reliable data stream processing with OSIRIS-SE. In G. Vossen, F. Leymann, P.C. Lockemann and W. Stucky (eds.), *BTW 2005. Datenbanksysteme in Business, Technologie und Web. Tagungsband der 11. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS). Karlsruhe (Germany), 2-4 March 2005*. Lecture Notes in Informatics (LNI) 65. Bonn: GI. 405-414.

[3] Brettlecker G., Schek H.-J. and Schuldt H. 2004. Information management infrastructure for telemonitoring in healthcare. In: In E. Ammenwerth, W. Gaus, R. Haux, C. Lovis, K.P. Pfeiffer, B. Tilg and H.E. Wichmann (eds.), *Tagungsband der 49. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS 2004). Innsbruck (Austria), 26-30 September 2004*. Niebüll: Videel Verlag OHG.

[4] Brettlecker G., Schuldt H. and Schatz R. 2004. Hyperdatabases for peer-to-peer data stream processing. In *Proceedings of the IEEE International Conference on Web Services (ICWS 2004). San Diego (USA), 6-9 June 2004*. Washington: IEEE Computer Society. 358-366.

[5] Dittrich J.-P., Fischer P. M. and Kossmann D. 2005. AGILE: adaptive indexing for context-aware information filters. In F. Ozcan (ed.), *Proceedings of the ACM SIGMOD International Conference on Management of Data. Baltimore (USA), 14-16 June 2005*. New York: ACM Press. 215-226.

[6] Florescu D., Grünhagen A., Kossmann D. 2003. XL: a platform for web services. In CIDR 2003, First Biennial Conference on Innovative Data Systems Research. Asilomar (USA), 5-8 January 2003.

[7] Fischer P. M. and Kossmann D. 2005. Batched processing for information filters. In *Proceedings of the 21st International Conference on Data Engineering (ICDE 2005). Tokyo (Japan), 5-8 April 2005*. Washington: IEEE Computer Society. 902-913.

[8] Wurz M., Brettlecker G.and H. Schuldt 2004. Data stream management and digital library processes on top of a hyperdatabase and grid infrastructure. In M. Agosti, H.-J. Schek and C. Türker (eds.), *Digital Library Architectures: Peer-to-Peer, Grid, and Service-Orientation. Pre-Proceedings of the 6th Thematic Workshop of the EU Network of Excellence DELOS. S. Margherita di Pula (Italy), 24-25 June 2004*. Padua: Edizioni Libreria Progetto. 37-48.

DelosDLMS: Global Prototype Development[4]

Hans Joerg Schek, University of Konstanz, schek@inf.ethz.ch
Heiko Schuldt, University of Basel, heiko.schuldt@unibas.ch

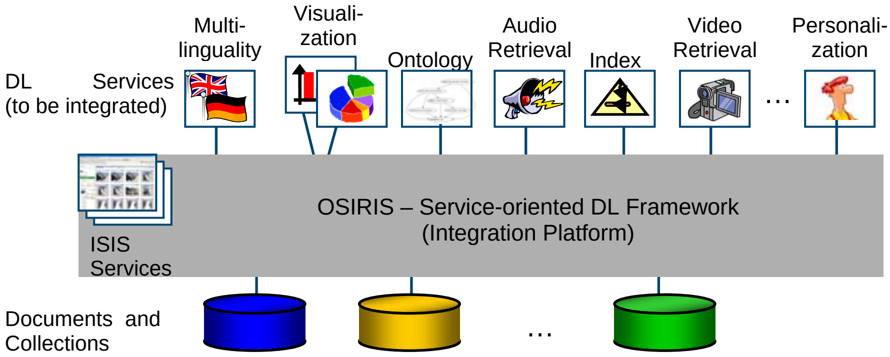
Introduction

In the first two years of the EC-funded network of excellence DELOS (A Network of Excellence on Digital Libraries), work has mainly focused on improving digital libraries (DLs) by developing independent, powerful and highly sophisticated prototype systems. The overall goal of the DelosDLMS is the implementation of a prototype of a next-generation digital library management system. This system combines text and audio-visual searching, offers personalized browsing using new information visualization and relevance feedback tools, allows retrieved information to be annotated and processed, integrates and processes sensor data streams, and finally, from a systems engineering point of view, is easily configured and adapted while being reliable and scalable. The prototype will be built by integrating digital library functionality provided by the DELOS partners into the OSIRIS/ISIS platform, a middleware environment developed by ETH Zürich and now being extended at the University of Basel. The result of the integration – that is, the middleware infrastructure together with all the advanced DL functionality – will constitute the DelosDLMS.

Research Objectives

A central task in the second phase of DELOS is the development of a global prototype. The objective is to build a joint prototype for the future Digital Library Management System that makes available results from many groups in DELOS. This will be based on the OSIRIS/ISIS middleware, the development of which began at ETH Zürich for ETHWorld, the virtual campus of ETH. It was further developed for data streams and for medical objects at UMIT, and is currently being extended at the University of Basel. The OSIRIS middleware (Open Service Infrastructure for Reliable and Integrated process Support) supports programming-in-the-large; i.e., the combination of arbitrary application services into so-called processes. This is realized by a set of generic (application-independent) services that include the registration of services and processes, interfaces for application development, an engine for decentralized execution of processes, and services for load balancing. In addition, it features reliable execution by applying advanced database concepts – essentially for failure handling and concurrency control – at the level of processes. ISIS (Interactive SImilarity Search) is a set of DL services that have been developed on the basis of the OSIRIS middleware. ISIS includes a sophisticated index structure (VA-file) for similarity searching, which is particularly well suited for high-dimensional vector spaces. Furthermore, in terms of Digital Library functions, ISIS features rudimentary support for textual and content-based audiovisual searching. It also provides basic support for relevance feedback and visualization.

With the DelosDLMS, existing ISIS services will be significantly enriched by other specialized DL services that have been developed within the DELOS network. This will be achieved by integrating these services into the OSIRIS infrastructure, thereby combining them with other ISIS and non-ISIS services into advanced, process-based DL applications.



[4] This joint research is supported by the EU in the 6th Framework Programme within the Network of Excellence DELOS (contract no. G038-507618).

Expected Results

The plan for the DelosDLMS includes the upgrading of existing ISIS components and services and the integration of new functionalities. The final product will support multi-object multi-feature queries over collections of different media types. Personalized browsing and information access, relevance feedback and object annotation will also be considered. Since information (for instance in e-Science Digital Library applications) increasingly originates from software or hardware sensors, sensor data stream processing will also be integrated in the DelosDLMS. Essentially, all this DL functionality will be made available by means of services. The challenge of DelosDLMS is therefore to provide a scalable and reliable infrastructure where these services can be plugged in and used as building blocks.

Two alternatives exist for integrating services with OSIRIS. First, there are tightly coupled services, which are tightly integrated into the OSIRIS runtime infrastructure. Advanced failure handling and load balancing are among the main advantages of this arrangement. In terms of failure handling, compensating services can be registered which are automatically invoked in case of failures. In terms of load balancing, ORISIS can automatically choose the node carrying the lightest load to invoke a service that is deployed several times. This is particularly important for computationally expensive services like feature extraction.

Second, services can be loosely coupled with OSIRIS, meaning that services are described and invoked by standard Web service interfaces (SOAP and WSDL). This reduces the effort needed for integration but does not provide the benefits of tight coupling.

In May 2006, a ‘call for services’ has been issued to both members and non-members of DELOS [Del 06]. The goal is to identify services which are best suited for integration into the DelosDLMS. For the first version of the integrated DelosDLMS prototype, services will in most cases be loosely coupled. The final version will then support a higher degree of reliability by tightly coupling as many services as possible. The areas from which the services to be integrated are taken include (but not limited to): sophisticated term extraction from text, text indexing and collection management, annotation services, reliable sensor data management, multimedia indexing, automatic search process generation and personalization services, image feature extraction, 3D shape recognition, special indexing techniques for video retrieval, audio feature extraction and audio retrieval, advanced visualization services and visual relevance feedback, self-organising map visualization, services for user interface generation and for interface functionality, active paper (linking digital information and paper), ontology services and natural language access, preservation services, services for multi-lingual access.

This list will be extended and revised after the evaluation of the call for services and during the actual integration work. Nonetheless, it highlights examples of building blocks that will be considered for DelosDLMS.

Project Organization

The DelosDLMS integration task is coordinated by the DELOS partner University of Konstanz and University of Basel. The actual integration activities are done in close collaboration with the respective DELOS (and non-DELOS) partner who are the providers of services to be integrated.

References

[1] DelosDLMS: Call for Services. http://dbis.cs.unibas.ch/delos_website/jpa2/ CallServicesDelosDLMS.txt

[2] Schuler Ch., Schuldt H., Türker C., Weber R. and H.-J. Schek 2005. Peer-to-Peer execution of (transactional) processes. In *International Journal of Cooperative Information Systems* 14(4): 377-405.

[3] Schuler Ch., Türker C., Schek H.-J., Weber R. and Schuldt H. 2006. Scalable Peer-to-Peer process management. In *International Journal of Business Process Integration and Management* 1(2): 129-142.

[4] Schuler Ch., Weber R., Schuldt H., Schek H.-J. 2004. Scalable Peer-to-Peer process management – The OSIRIS approach. In *Proceedings of the IEEE International Conference on Web Services (ICWS 2004). San Diego (USA), 6-9 June 2004*. Washington: IEEE Computer Society. 26-34.

[5] Weber R., Schuler C., Neukomm P., Schuldt H. and Schek H.-J. 2003. WebService composition with O’Grape and OSIRIS. In J.Ch. Freytag, P.C. Lockemann, S. Abiteboul, M. J. Carey, P.G. Selinger and A. Heuer (eds.), *Proceedings of 29th International Conference on Very Large Data Bases (VLDB 2003). Berlin (Germany), 9-12 September 2003*. San Francisco: Morgan Kaufmann. 1081-1084.

Information Access and Personalization

Cluster objectives

Information stored in digital libraries needs to be accessed, integrated and individualized for any user anytime and anywhere in possibly multiple comprehensive and efficient ways. Within Delos, Information Access in Digital Libraries is studied from three different aspects.

Information Access: interaction with a single information provider. Information stored in a source comes in different types and formats, each one with its own characteristics and particularities. Organization of data within an individual source and efficient and effective search are the key issues and are actually highly interrelated to each other. Different approaches exist but there is a general trend towards richer representations and languages both at the structural and at the semantic level.
Information Integration: interaction with multiple information providers. Integrated access of different sources presents specialized problems due to information heterogeneity, redundancy etc. Issues such as source selection and results fusion must be considered under different possible settings. Data provenance is often crucial to the trust that is placed in data; hence it should be managed based on sound formulation.
Personalization: customization of information and interaction to user. Different users have different characteristics and preferences concerning the information they are interested in seeing when accessing a digital library. Even users sharing a common information need may expect different results, different functionality or different interface. Moreover, the relevant contents and interface of a digital library may be dependent on other factors as well, e.g. device or network-specific.

The cluster's objectives with respect to the aforementioned aspects are the following:

- Promotion of knowledge about available practices in the fields of information access and personalization in digital libraries is the first goal pursued. This will lead to a uniform understanding of problems among researchers.
- Construction of a common, comprehensive framework for information access and personalization approaches is essential. This framework is intended to serve as a reference point for the DL area and to stimulate research.
- Promotion of research on new information access and personalization models and methodologies.

Cluster activities

The cluster's activities with respect to information access, integration and personalization are very coarsely organized into the following categories:

- Collection, study, and comparison of models, languages and algorithms for data, metadata, and queries with respect to information access and integration
- Collection, study, and comparison of user-profile models and various forms of content and interaction personalization
- Integration of the most effective approaches to information access, integration, and personalization and derivation of new ones
- Development of toolkits and systems for purposes of re-use and demonstration of proposed methods and models

Cluster coordinator

Yannis Ioannidis, National and Kapodistrian University of Athens, Greece, yannis@di.uoa.gr

Advanced Access Structures for Complex Similarity Measures

Giuseppe Amato, ISTI-CNR, Giuseppe.Amato@isti.cnr.it
Benjamin Bustos, University of Konstanz, bustos@dbvis.inf.uni-konstanz.de,
Daniel Keim, University of Konstanz, keim@informatik.uni-konstanz.de
Tobias Schreck, University of Konstanz, schreck@dbvis.inf.uni-konstanz.de,
Pavel Zezula, MUNI, Zezula@fi.muni.cz

Research Problem and Objectives

Similarity search in large collections of media objects is typically obtained relying on some distance functions that compute the (dis)similarity between two objects. Distance functions are domain specific and may either operate on the objects themselves or on pre-computed feature vectors that express certain characteristics through numeric values. Many indexing proposals exist to support feature-based query processing by means of distance measures which compare features represented in vector spaces [1] and metric spaces [2].

The objective of this task is to refine and extend access structures for enhanced distributed search and increased search effectiveness. In particular, new directions will be explored to build access methods for image similarity search, such as access methods for region-based image retrieval. Image segmentation, clustering, and feature weighting will be exploited to index images in a compact way. Furthermore, new structures will be developed to achieve scalable but also tuneable performance of similarity search. All these will be complemented by joint experimentation among partners.

Work done

Various results have been obtained so far in the activities carried out in this task. These results have given fruitful hints on how to improve current techniques and suggested new interesting directions to be investigated.

An efficient indexing method was proposed that adopts quantization-based approximation techniques, pivot-based clustering, and a novel principle of self-refinement to yield substantial cost reduction in similarity search atop metric distance measures. Both an indexing and a retrieval component for ranked queries were implemented to support content-based image retrieval using complex colour matching algorithms. That implementation was later refined for straightforward intra-query parallelism that employs a simple communication protocol to speed up searches by distributing sub queries across multiple nodes processing distinct index chunks. Other modifications were introduced to cope with inter-query parallelism that constraints main memory consumption of query processing [8], [9], [10].

A pivot-based index structure, specially designed to support similarity queries with dynamically weighed combinations of feature vectors, has been developed [11]. This index consists of a set of pivot-based indices, one for each feature vector, which can be used to compute the combined pivot table (i.e., the pivot-based index for the combination of feature vectors) at query time, when the weights for the dynamic combination are known.

Techniques for digital library oriented similarity search techniques have been developed. Specifically, given the current standardisation initiatives, consider for instance the MPEG-7 initiative, of encoding metadata (including low level features and conceptual information) as XML documents, we have developed techniques for processing combined exact and similarity match queries on XML encoded metadata. These techniques were deployed in the implementation of a native XML database for digital library applications [4], which is a core component of the MILOS multimedia content management system for digital libraries [3], [5].

A new scalable and distributed similarity search structure called M-CAN [7]was also developed. The main idea is to transform the metric search problem into a low dimensional problem and apply the key-word search Peer to Peer communication protocol called CAN. At the moment we are working on an extensive experimental comparison of different scalable and distributed similarity search structures.

Also a deep survey of the state of the art on techniques for similarity search of data represented in metric spaces was prepared. This high quality survey has been published as a book by Springer [6].

New research directions

Research directions and joint activities are directed towards integrating, disseminating, and refining the aforementioned results.

By conducting joint experiments (1) relating to common efficiency and effectiveness measures and (2) including an exchange of implementations and test data sets, partners will provide a verifiable assessment of their achievements.

More effort will be spent on refining and extending access structures for enhanced distributed search and increased search effectiveness. In detail, we plan to introduce asynchronous communication between processing nodes that exchanges intermediate query results (1) to accelerate the overall pruning of the search space, (2) to early shut down query processing on individual nodes to free resources for concurrent queries, (3) to increase robustness and failure tolerance where query workloads may be dynamically re-scheduled, and (4) to permit increased adaptiveness to system configuration and current workload.

We plan to investigate new directions to build access methods for image similarity search. Specifically we will study access methods for region based image retrieval, by adapting the inverted files technique to deal with this type of data. We will exploit techniques for image segmentation, clustering algorithms, and feature weighting techniques to index images in a compact way, which allow the efficient and effective use of inverted files for image similarity searching.

We plan to capitalize on previous results, both experimental and theoretical, and develop new structures with the main objective to achieve not only scalable but also tuneable performance of similarity search.

Expected results

The task is expected to yield a number of versatile, efficient and effective access structures for non-standard similarity searches in various domains. Particular effort will be spent on scalability (w.r.t. query workload parallelization) which is of crucial importance to cope with the intrinsic expensiveness of distance-based similarity search. Emphasis will also be put on good effectiveness of query outcomes by supporting complex query modes, relevance feedback mechanisms, and sophisticated matching algorithms acting atop multiple feature types.

This task provides a crucial contribution to permitting the DELOS vision come into being. Efficient and effective support for non-traditional search operations in large data stocks of multimedia documents is a distinct prerequisite for future digital libraries where appropriate means of finding documents in a timely manner must be provided. In coping with the similarity search paradigm, ranked content-based searches in multimedia document collections are specifically addressed.

References

[1] Böhm C., Berchtold S. and Keim D.A. 2001. Searching in high-dimensional spaces – Index structures for improving the performance of multimedia databases. In *ACM Computing Surveys* 33(3): 322-373.

[2] Chavez E., Navarro G., Baeza-Yates R. and Marroquin J.L. 2001. Searching in metric spaces. In *ACM Computing Surveys* 33(3): 273-321.

[3] Amato G., Gennaro C., Rabitti F. and Savino P. 2004. MILOS: a multimedia content management system for digital library applications. In R. Heery and L. Lyon (eds.), *Research and Advanced Technology for Digital Libraries. Proceedings of the 8th European Conference (ECDL 2004). Bath (UK), 12-17 September 2004*. Lecture Notes in Computer Science 3232. Berlin-Heidelberg: Springer. 14-25.

[4] Amato G. and Debole F. 2005. A native XML database supporting approximate match search. In A. Rauber, S. Christodoulakis, A. Min Tjoa (eds.), *Research and Advanced Technology for Digital Libraries. Proceedings of the 9th European Conference on Digital Libraries (ECDL 2005). Vienna (Austria), 22 September 2005*. Lecture Notes in Computer Science 3652. Berlin-Heidelberg: Springer. 69-80.

[5] Amato G., Gennaro C., Rabitti F. and Savino P. 2005. Functionalities of a content management system specialised for digital library applications. In T. Catarci, S. Christodoulakis and A. Del Bimbo (eds.), *AVIVDiLib'05. Proceedings of the 7th International Workshop of EU Network of Excellence on Audio-Visual Content and Information Visualization in Digital Libraries. Cortona (Italy), 4-6 May 2005*. DELOS Network of Excellence. 47-56.

[6] Zezula P., Amato G., Dohnal V. and Batko, M. 2006. *Similarity Search. The Metric Space Approach*. Advances in Database Systems 32. Berlin-Heidelberg: Springer.

[7] Falchi F., Gennaro C. and Zezula P. 2005. A content addressable network for similarity search in metric spaces. 3rd International Workshop on Databases, Information Systems, and Peer-to-Peer Computing (DBISP2P 2005). Toronto (Canada), 29-30 August 2004.

[8] Schmitt I. and Balko S. 2006. Filter ranking in high-dimensional space. In Data and Knowledge Engineering 56(3): 245-286.

[9] Mlivoncic M., Schuler Ch., Türker C. and Balko S. 2005. A service-oriented grid infrastructure for multimedia management and search. In C. Türker, M. Agosti and H.-J. Schek (eds.), *Peer-to-Peer, Grid, and Service-Orientation in Digital Library Architectures. Revised Selected Papers from the 6th Thematic Workshop of the EU Network of Excellence DELOS. S. Margherita di Pula (Italy), 24-25 June 2004*. Lecture Notes in Computer Science 3664. Berlin-Heidelberg: Springer. 167-187.

[10] Balko S. 2005. High-dimensional Indexing - Formal Foundations and Novel Approaches. Lecture Notes in Informatics D-5.

[11] Bustos B., Keim D. and Schreck T. 2005.. A pivot-based index structure for combination of feature vectors. In H. Haddad, L.M. Liebrock, A. Omicini and R.L. Wainwright (eds.), *Proceedings of the ACM Symposium on Applied Computing (SAC 2005)*. Santa Fe (USA), 13-17 March 2005. New York: ACM Press. 1180-1184.

Application of the Peer-to-Peer Paradigm in Digital Libraries

Vassilis Christophides, ICS, FORTH, christop@ics.forth.gr
Theodore Dalamangas, NTUA, dalamag@dblab.ece.ntua.gr
Timos Sellis, NTUA, Athens, timos@dblab.ece.ntua.gr

Summary

A Digital Library (DL) can be thought of as an interconnected network of individual library nodes. These nodes can be part of the same physical network (*e.g.*, the library of a single institution), or they can be part of a larger, distributed network (*e.g.*, a common gateway to institutional libraries). In both scenarios, however, the key idea is to provide transparent access to a single virtual network; in other words, we want to *abstract* the network structure so the user of the DL is presented with a single view to the DL's data.

On the other hand, Peer-to-Peer (P2P) systems are distributed systems in which no centralised authority exists and all peers are considered to provide the same functionality. The key premise of a P2P system is providing a decentralised implementation of a dictionary interface so that any peer in the system can efficiently locate any other peer responsible for a given data item. The semantics of the data stored in a P2P system are handled by the application, rather than the storage layer.

It seems natural to combine the two into a single framework. In this proposal we will present our approach to extending Peer-to-Peer technology and tailoring it to the Digital Library domain. In a nutshell, the P2P paradigm will act as the transparent storage layer of a largely distributed DL.

Our vision for a P2P DL involves data management in a dynamic network of autonomously managed peers. Autonomy in this context refers to the following objectives:

Joining and leaving the library. Each peer in the system chooses itself when to join and/or leave the library network. The system should gracefully adapt to these joins and departures without any global structural knowledge.

Data management. A peer is responsible for managing its own local data, as well as maintaining information about data residing at other peers.

Query processing. Given a query posed at any peer of the system, the system must first identify the peers capable of answering the query and then optimise and evaluate the query across those peers.

To facilitate these objectives, we propose an architecture based on *schema mappings* and *query reformulation*. The high-level description of such an architecture is presented in Figure 1. Each peer in the system has a data model its own data conform to. However, it exports a description of its schema, mapped to a *common data model* across all peers of the system; for our purposes, this common data model is XML or RDF (see *e.g.*, [1, 5]). We refer to the exported schema as the *local schema*} and it is effectively the data that each peer is capable of “serving” to other peers in the system[5].

Once peers have exported their schemata, the next step is combining this information to allow seamless querying across the peers of the entire system. This will be achieved through *schema mappings*. More specifically, mappings are ways of maintaining schema correspondences between the various peers of the system. These mappings allow a peer to translate a locally posed query to the schema exported by remote peers, so these peers can contribute to the answer. The key issue here is identifying these mappings in the first place and keeping them up to date as the system evolves. In addition, mappings are not only used to translate queries, but also to identify the peers that can be used to answer a specific query. Therefore, the mappings act as a decentralised index that is used to *route* data retrieval requests to the peers capable of serving them. This identification and translation process is termed *query reformulation* (see *e.g.*, [2, 4]).

[5] An interesting aspect here is having the peers export different schemata to different peers; this presents a way for the peers of the system to implement their own security models.

After a query has been reformulated it needs to be evaluated across the participating peers. The query needs to be optimised so all individual queries posed to the peers are efficiently scheduled and their results appropriately combined. This calls for novel techniques for *distributed query optimization and evaluation* in the context of a decentralised P2P system (see *e.g.*, [3, 6, 7]). For instance, an important aspect is making sure the number of visits to each participating peers is minimised; at the same time, individual queries need to be scheduled in such a way so as intermediate result sizes are minimised. Another interesting problem stems from the sheer scale of a largely distributed system. The number of evaluation plans considered is even larger than in a traditional query processing system. Already a hard problem for centralised systems, the additional parameter of discovering the relevant data through mappings and incrementally having to reformulate the results leads to an even greater search space explosion. Furthermore, there may be competitive metrics: for example, one user may be more interested in the quantity of returned results (*i.e.*, taking as many mappings into account as possible) while another may be interested in response time (*i.e.*, receiving an answer fast, without the answer necessarily being of the highest quality, for some metric of quality). Scenarios like the one described give way to modeling query processing as a multi-objective optimisation problem. Clearly, this calls for novel query optimisation and processing techniques.

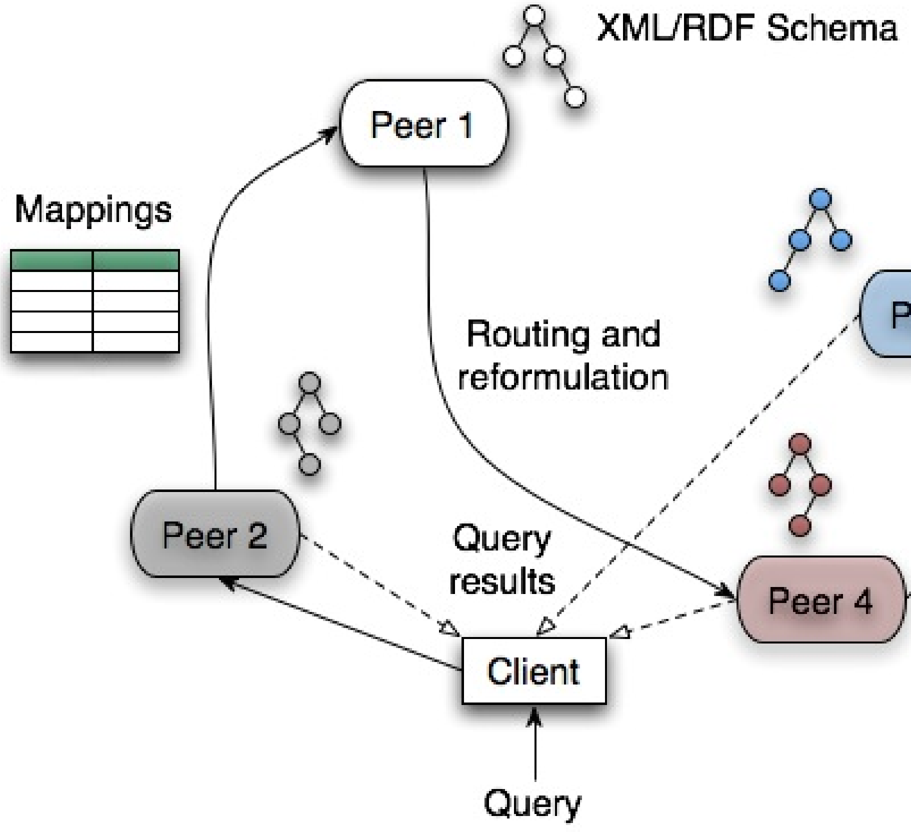


Figure 1: The architecture of a P2P DL

In light of the proposed architecture of a P2P DL, and the problems described earlier we have identified three main axes to our research, summarised as follows:

Logical foundations of query reformulation. We will develop a theory modeling query reformulation in largely distributed systems. More specifically, we will focus on the logical foundations behind query mappings and how these foundations can be used to declaratively reason about queries posed across multiple sites. Based on the logical formalisms, we will develop procedural and optimisable ways of evaluating queries in a distributed way.

Routing indexes. We will develop primitives for storing and maintaining mappings across the peers of a distributed system, to facilitate efficient access to the information stored at each peer. These indexes will aid in *(a)* identifying the peers relevant to a particular query, and *(b)* routing a complicated data retrieval request across the peers of the entire system. Moreover, we will address the problem of maintaining these indexes as the system evolves, *i.e.*, as peers join and leave the network and/or as the data exported by each peer changes.

Distributed query optimization and evaluation. We will identify new distributed query processing paradigms that are directly applicable to the P2P setting. More specifically, we will identify the appropriate metrics that characterise performance in such an environment, and provide ways of employing these metrics in optimizing distributed queries in a decentralised manner. Moreover, we will introduce query evaluation algorithms and techniques for efficiently executing queries across the peers of the system; these techniques will be in accordance to the newly identified performance metrics.

Note that research along these lines is not strictly confined to DLs, but also advances the state of the art in distributed query processing in general.

Workplan

The three research axes we have identified, namely logical foundations of query reformulation, routing indexes and query processing and optimisation objectives, form the basic workplan for the duration of this project.

More specifically, we plan to begin with the formalisation of a P2P DL and focus on the declarative and procedural abstractions that will allow us to further tackle the problem from a theoretical point of view. Based on these outcomes, we will refine our approach to building scalable and decentralised routing indexes that will aid in implementing the aforementioned abstractions.

In more concrete terms, we plan to first develop a simulator of our proposal. The objective is to have a highly scalable simulator of a real P2P DL system, that will allow as to identify potential bottlenecks of such a system. We expect the formalisation of the framework and implementation of the simulator to last for twelve man-months, but those two objectives will not be proceeding in parallel for the full duration; formalisation will be tackled during the first three man-months and once the basic building blocks are in place, implementation of the simulator will commence and further refinement of the formal model will proceed in parallel for the remaining nine man-months.

Once the simulator is in place, we will start experimenting with various network settings and workload scenarios to explore the boundaries of our approach. We expect this experimentation to last for a further three man-months. Based on the simulation results, we will start developing a concrete implementation of a prototype P2P DL system, aiming for a real deployment. This development is expected to last for a further nine man-months, followed by additional experimentation to identify possible problematic scenarios that have not been pointed out during simulation.

References

[1] Kaoudi Z., Dalamagas T. and Sellis T.K. 2005. RDF-Sculpt: managing RDF Schemas under Set-like semantics. In A. Gómez-Pérez and J. Euzenat (eds.), The Semantic Web: Research and Applications. Proceedings of the 2nd European Semantic Web Conference (ESWC 2005). Heraklion (Greece), 29 May - 1 June 2005. Lecture Notes in Computer Science 3532. Berlin-Heidelberg: Springer. 123–137.

[2] Kokkinidis G. and Christophides V. 2004. Semantic query routing and processing in P2P database systems: the ICS-FORTH SQPeer middleware. In W. Lindner, M. Mesiti, C. Türker, Y. Tzitzikas and A. Vakali (eds.), *Current Trends in Database Technology. Revised Selected Papers from the EDBT 2004 Workshops, EDBT 2004 Workshops PhD, DataX, PIM, P2P&DB and ClustWeb. Heraklion (Greece), 14-18 March 2004.* Lecture Notes in Computer Science 3268 Berlin-Heidelberg: Springer. 486-495.

[3] Kokkinidis G., Christophides V., Dalamagas T. and Viglas S. 2005. Query processing in P2P database management systems: a state-of-the-art. Technical Report, Apr 2005. IST-FP6, DELOS Network of Excellence, NoE G038-507618.

[4] Sidirourgos L., Kokkinidis G. and Dalamagas T. 2005. Efficient query routing in RDF/S schema-based P2P systems. In *4th Hellenic Data Management Symposium (HDMS 2005). Athens (Greece), 25-26 August 2005.*

[5] Spyropoulou E. and Dalamagas T. forthcoming. SDQNET: a platform for semantic query processing in loosely coupled data sources. In *Proceedings of the 10th East European Conference on Advances in Databases and Information Systems (ADBIS 2006). Thessaloniki (Greece), 3-7 September 2006.*

[6] Viglas, S.D. 2004. Pyragrid: bringing peer-to-peer and grid architectures together. In M Agosti, H.-J. Schek and C. Türker (eds.), *Digital Library Architectures: Peer-to-Peer, Grid, and Service-Orientation. Pre-proceedings of the 6th Thematic Workshop of the EU Network of Excellence DELOS. S. Margherita di Pula (Italy), 24-25 June 2004*. Padua: Edizioni Libreria Progetto. 1-12.

[7] Viglas S.D. 2006. *Distributed File Structures in a Peer-to-Peer Environment*. Technical Report. School of Informatics, University of Edinburgh.

Personalized Query Routing in Peer-to-Peer Federations of Digital Libraries

Matthias Bender, Max-Planck Institute for Informatics, mbender@mpi-inf.mpg.de
Norbert Fuhr, University of Duisburg, fuhr@uni-duisburg.de
Yannis Ioannidis, National and Kapodistrian University of Athens, yannis@di.uoa.gr
Donald Kossmann, Swiss Federal Institute of Technology (ETH) Zürich, donald.kossmann@inf.ethz.ch
Hans Joerg Schek, University of Konstanz, schek@inf.ethz.ch
Gerhard Weikum, Max-Planck Institute for Informatics, weikum@mpi-sb.mpg.de
Pavel Zezula, MUNI, Zezula@fi.muni.cz
Christian Zimmer, Max-Planck Institute for Informatics, czimmer@mpi-inf.mpg.de

Research Problem and Objectives

The peer-to-peer (P2P) paradigm [SW05] is a promising approach for coping with dynamically evolving federations of loosely coupled digital libraries where user agents with powerful personalized tools may participate besides the libraries as peers, too. The advantages include scalability, failure robustness, and reduced vulnerability to information manipulation or attacks. The participating peers are autonomous and collaborate on behalf of user requests at their discretion. A user peer applies its local profile to pose appropriate queries to the most suitable library peers for the given information demand. All-dominant is the decision about which libraries should receive a query. In the literature, the decision is known as *query routing* or *database selection* or *resource selection*. In the area of distributed IR and metasearch engines [LC03, NF03, MYL02], database selection has been intensively studied, but the previous work assumed a static architecture with a proportionally small number of peers. Dynamic aspects have not been considered at all, however the large scale and the high dynamics of a P2P system make new demands on the system architecture. Each peer can share the users' access patterns, query logs, and further user-behavior information with other peers; this provides great opportunities for advanced personalization of information selection and query execution, leading to better search result quality.

The research work about personalized query routing considers user preferences, query logs, click streams, and further long-term information about user behavior reflecting the user's information needs and biases [Be04, KI05a, KI05b, LW06]. In the context of personalized query routing, the information requests include simple keyword queries and structured queries (SQL or XQuery) and may also consider entire sessions of searching, browsing, annotation, classification, and refinement steps. Another idea considers the expansion or rewriting of a user query based on the local profile or local document collection. This way, a query can emphasis personalized preferences. The work specifically focuses on how to exploit the above kind of user-behavior information for the purpose of query routing in the distributed P2P federation.

Approach and Results

In this task, most of the research work is conceptual in nature: planning and designing models and strategies for profiling and personalized query routing. A critical issue to be considered is the evaluation of strategies. A comprehensive experimental evaluation is beyond the scope of this work, but the conceptual work should already take this later stage into account by defining an appropriate experimental setup.

A key aspect is to model and represent the user’s individual interests and preferences in a compact *profile*. [KI05a, KI05b] have developed an expressive and versatile profile model that allows to incorporate constraints in a flexible manner. Its benefit for personalized search has been demonstrated on structured databases; its generalization to semi structured data in digital libraries is being pursued. Peer-specific profiles are also beneficial for query result summarization, as demonstrated in [KS06]. Finally, when peers have their own data collections (e.g., bookmarks, individual working sets, annotated documents, etc.), overlap awareness can improve query routing and collaborative search by estimating the novelty of information that is obtainable from different users and digital libraries [Be05a, Mi06]. A tutorial on profile models for personalized search has been given at the VLDB conference [KI05c].

By utilizing *query logs* and the click behaviour of users, peers can identify strongly correlated terms and improve their query routing by extending the directory knowledge. Experimental tests given in [Be06a] showed that this approach leads to significant improvements in query result quality. The same information can also be leveraged for enhanced authority ranking [LW06], which in turn can lead to better ways of organizing peers in a semantic overlay network [PMW05]. In all these approaches, statistical synopses need to be constructed and disseminated among peers or posted to a decentralized directory [Be05a, Mi06]. For efficiency, it may be desirable to restrict the synopses to the topics and content features that are truly specific for a given peer, reflecting the individual strengths in the peer's profile [Be06b, Po06]. The experiments in [Be06b] have shown that this yields a very effective and low-overhead method for personalized query routing.

Another research direction studies the P2P paradigm for *publish-subscribe* applications where queries are continuous queries according to user profiles. In this information filtering (or continuous querying) scenario, research has the focus on specific kinds of queries related to time and location of events (e.g. news or medical information) over semi structured data collections. Here, in contrast to personalized search with one-time queries, the temporal evolution and anticipated future behaviour of user subscriptions and the publishing patterns of digital libraries have to be considered. [Di05] has developed an efficient indexing method for such settings with particularly good adaptation to evolving situations. Process management in a P2P network is another important key aspect for satisfying more complex information demands; efficient protocols with strong guarantees are presented in [Ha05, Sch05].

For experimental evaluation, we have developed a P2P prototype testbed which can also easily be installed at the sites of the partners participating in this work. This prototype system, coined Minerva [Be05a, Be05b], can be used for testing and evaluating different approaches concerning personalized query routing. An important aspect is that the testbed is open for easy adoption of alternative strategies. As a first step, Minerva has been coupled with M-Chord [NZ06, Ze05] to support edit-distance-based spelling corrections in combination for keyword search. A comprehensive scalability study of four distributed similarity search structures can be found in [Ba06]. [NF06] has compared different architectures for query routing, including the decision-theoretic framework and its possible adaptation to a personalized setting. Issues of experimental evaluation have also been discussed at an international workshop on P2P IR [No05].

References

[1] Batko M., Novak D., Falchi F. and Zezula P. forthcoming. On scalability of the similarity search in the world of peers. In *Proceedings of the 1st IEEE International Conference on Scalable Information Systems (INFOSCALE 2006). Hong Kong (China), 30 May – 1 June 2006.*

[2] Bender M., Michel S., Zimmer Ch. and Weikum G. 2004. Bookmark-driven query routing in peer-to-peer web search. In J. Callan, N. Fuhr and W. Nejdl (eds.), *Proceedings of the SIGIR Workshop on Peer-to-Peer Information Retrieval, 27th Annual International ACM SIGIR Conference. Sheffield (UK), 25-29 July 2004*. New York: ACM Press.

[3] Bender M., Michel S., Triantafillou P., Weikum G. and Zimmer Ch. 2005. Improving collection selection with overlap awareness in P2P search engines. In R.A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat and J. Tait (eds.), *SIGIR 2005. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Salvador (Brazil), 15-19 August 2005*. New York: ACM Press. 67-74.

[4] Bender M., Michel S., Triantafillou P., Weikum G. and Zimmer Ch. 2005. MINERVA: collaborative P2P search. In K. Böhm, Ch.S. Jensen, L.M. Haas, M.L. Kersten, P.-Å. Larson and B.C. Ooi (eds.), *Proceedings of the 31st International Conference on Very Large Data Bases (VLDB 2005). Trondheim (Norway), 30 August - 2 September 2005*. New York: ACM Press. 1263-1266.

[5] Bender M., Michel S., Triantafillou P., Weikum G. and Zimmer C. 2006b. P2P content search: give the Web back to the people. In *Proceedings of the 5th International Workshop on Peer-to-Peer Systems (IPTPS 2006). Santa Barbara (USA), 27-28 February 2006*.

[6] Bender M., Michel S. and Weikum G. 2006. P2P directories for distributed web search: from each according to his ability, to each according to his needs. In *Proceedings of the International Workshop on Web Information Retrieval and Integration (WIRI). Atlanta (USA), 3 April 2006.*

[7] Dittrich J.-P., Fischer P. M. and Kossmann D. 2005. AGILE: Adaptive indexing for context-aware information filters. In F. Ozcan (ed.), *Proceedings of the ACM SIGMOD International Conference on Management of Data. Baltimore (USA), 14-16 June 2005*. New York: ACM Press. 215-226.

[8] Haller K., Schuldt H. and Türker C. 2005. Decentralized coordination of transactional processes in peer-to-peer environments. In O. Herzog, H.-J. Schek, N. Fuhr, A. Chowdhury and W. Teiken (eds.), *Proceedings of the*

2005 ACM CIKM International Conference on Information and Knowledge Management. Bremen (Germany), 31 October 5 November 2005. New York: ACM Press. 28-35.

[9] Koutrika G. and Y. Ioannidis 2005. Personalized queries under a generalized preference model. In *Proceedings of the 21st International Conference on Data Engineering (ICDE 2005). Tokyo (Japan), 5-8 April 2005*. Washington: IEEE Computer Society. 841-852.

[10] Koutrika G. and Y. Ioannidis 2005. Constrained optimalities in query personalization. In F. Ozcan (ed.), *Proceedings of the ACM SIGMOD International Conference on Management of Data. Baltimore (USA), 14-16 June 2005*. New York: ACM Press. 73-84.

Context-dependent Access to Digital Libraries

Yannis Ioannidis, National and Kapodistrian University of Athens, yannis@di.uoa.gr
Georgia Koutrika, National and Kapodistrian University of Athens, koutrika@di.uoa.gr
Timos Sellis, NTUA-ICCS , timos@dblab.ece.ntua.gr
Nicolas Spyratos, University of Paris-Sud XI, spyratos@lri.fr
Yannis Stavrakas, NTUA-ICCS, ys@dblab.ntua.gr

Research Problems

DELOS envisages Digital Libraries that would allow information to be accessed and used in a global environment, where implicit assumptions about data become less and less evident. Users with different backgrounds or viewpoints may interpret the same data in different ways. Moreover, the *interpretation* and *suitability* of data may depend on changing conditions, e.g. the current position of the user or the media he is using (laptop, mobile, PDA, etc.). To cater for such ambiguous situations the information provider needs to specify the *context* under which information becomes relevant. Conversely, information users can specify their own current context when requesting for data in order to denote the part that is relevant to their specific situation.

Task 2.9 aims to investigate how support for context-dependent data can be integrated into Digital Libraries. The goal is to use context as an abstraction mechanism that will allow Digital Libraries to provide users with information relevant to their situation, background, and preferences. The direction is twofold: (a) incorporate context as first class citizen into data management systems, and (b) use context to enhance the support for personalized access to Digital Libraries.

Specifically, in Task 2.9 we consider the following topics:

- Ways to express context.
- Models that allow to associate information with the relevant context.
- Inclusion of context into the process of data retrieval.
- Ways to use context for personalized information delivery.

Work done

So far, the results of Task 2.9 are as follows:

- Context-aware data: logical model and operations

We have defined the Context Relational model (CR model), a model that extends the relational model by incorporating the notion of context. We have also defined a number of basic, relational algebra-like operations on the CR model, which incorporate context. The interesting part of this approach is that context is treated as first-class citizen at the level of the database management system.

- Personalized information delivery

We have worked on user preference modeling for database queries, on integrating real time factors of the query context, such as response time requirements, into query personalization based on user preferences. Work has also been directed in personalization of keyword queries based on structured user profiles.

- Prototype implementation of a context-aware relational database

We have implemented a prototype system based on the CR model. The system allows to define “context relations”, that is, relations in which context is used to determine the structure and value of information entities. Data can be retrieved using a number of relational algebra-like operations that incorporate context.

- Storing context-aware data

We have looked into a number of alternative techniques for storing “context relations” of the CR model, using conventional relational tables. Different storing techniques influence the way information is subsequently retrieved. We have investigated how the choice of a storage technique affects the performance of context-aware operations.

Objectives

The major research directions for Task 2.9 in JPA3 will be the following:

- Context modelling

Continuation of the work done in the CR model, for incorporating context into relational database management systems: a context-aware query language is needed that will combine the operations we have defined. Also, evaluation of context-dependent queries, and investigation of access methods that take query context into account. Exploration of new formalisms for expressing context: so far context is expressed by assigning discrete values to user-defined variables, which is adequate for capturing user preferences. However there exist situations where this is not enough, for example when monitoring the position of a user.

- Context and personalization

Context can express user preferences. As shown in previous work by NTUA-ICCS and UPS-XI, a context-aware data management system can then apply application-independent processes to deliver personalized information to users. On the other hand, user preferences may themselves be context-dependent. We plan to look into both directions, and investigate (a) how a context-aware model can be used to support personalization in a uniform way at the data management system level, and (b) how the notion of context can enhance the personalization methods previously proposed by UoA.

- Dynamically definable context

So far we focused on a notion of context that is static and defined in advance. However, in many real world applications context can not be determined in advance, and changes over time. In other cases context can be inferred by user actions. For example, in an e-library a context navigation tree may act as a guide to the appropriate items of interest associated with the user's context The user's list of items of interest will be a possibly incomplete navigation tree which will assist the user in finding the data he/she is looking for.

- Cooperation with other DELOS Tasks

Other DELOS Tasks approach personalized information access from different directions. In JPA3 we plan to strengthen the interaction between Tasks involving context, integration of content, personal ontologies, smart browsing, relevance feedback, and query refinement.

References

[1] Theodorakis M., Analyti A., Constantopoulos P. and Spyratos N. 2001. A theory of contexts in information bases. In Information Systems Journal 19(4): 1-54.

[2] Stavrakas Y., Gergatsoulis M., Doulkeridis Ch. and Zafeiris V. 2004. Representing and querying histories of semistructured databases using multidimensional OEM. In *Information Systems* 29(6): 461-482.

[3] Koutrika G. and Y. Ioannidis 2005c. Personalized queries under a generalized preference model. In *Proceedings of the 21st International Conference on Data Engineering (ICDE 2005). Tokyo (Japan), 5-8 April 2005*. Washington: IEEE Computer Society. 841-852.

[4] Koutrika G. and Ioannidis Y. 2004. Personalization of queries in database systems. In *Proceedings of the 20th International Conference on Data Engineering (ICDE 2004). Boston (USA), 30 March - 2 April 2004*. Washington: IEEE Computer Society. 597-608.

[5] Roussos Y., Stavrakas Y., and Pavlaki V. 2005. Towards a context-aware relational model. In Proceedings of the Contextual Representation and Reasoning Workshop of the 5th International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT 2005). Paris (France), 5-8 July 2005.

[6] Akaishi M., Spyratos N. and Tanaka Y. 2003. Contextual search in large collections of information resources. In Y. Kiyoki, E. Kawaguchi, H. Jaakkola and H. Kangassalo (eds.), *Information Modelling and Knowledge Bases XV. Proceedings of the 13th European-Japanese Conference on Information Modelling and Knowledge Bases (EJC 2003). Kitakyushu (Japan), 3-6 June 2003*. Frontiers in Artificial Intelligence and Applications 105. Amsterdam: IOS Press. 295-302.

Modeling of User Preferences in Digital Libraries

Vassilis Christophides, ICS, FORTH, christop@ics.forth.gr
Yannis Ioannidis, National and Kapodistrian University of Athens, yannis@di.uoa.gr
Carlo Meghini, ISTI-CNR, carlo.meghini@isti.cnr.it
Nicolas Spyratos, University of Paris-Sud XI, spyratos@lri.fr

Task Content Description

As information becomes available in increasing amounts to growing numbers of users, the shift towards a more user-centered, or *personalized access* to information is becoming an increasingly important issue.

Personalized access can involve customization of the user interface or adaptation of the content to user's needs or *preferences*. This task addresses adaptation of content to user's preferences, and aims at the following objectives:

- study a formal framework for specifying user preferences
- enrich the digital library services with preference capabilities
- design algorithms for supporting preferences

Two main approaches to the study of preferences are considered, the *qualitative approach* and the *quantitative approach*. In the qualitative approach preferences are expressed by comparing different choices that is by using "preference relations" (eg. **"**I like A *better than* B"), while in the quantitative approach preferences are expressed by evaluating each individual choice that is by using "scores" (eg. "I like A *very much*", I like B *a little*"). This task aims at studying both approaches. The main results expected are as follows:

- formal framework for the definition of qualitative/quantitative preferences
- demonstrator toolkit

Two lines of research seem to be the natural evolution of the work conducted so far. The first concerns "tie-breaking" that is discrimination between documents that are ranked equally (i.e., they are equivalent in terms of user preferences). One possible approach is asking the user to express preferences in terms of additional document facets, for example in terms of content, in terms of language, in terms of format etc. (the more the facets, the higher the discrimination).

The second line of research concerns "consensus ranking" that is ranking that respects the preferences not of just one user but rather of a group of users. This kind of ranking is useful in the context of so called "social networks", where social groups are formed by individuals with similar preferences. A large body of research already exists in fields other than computer science (e.g. in operations research) from which our research could benefit.

References

[1] Spyratos N. and Meghini C. 2006. Preference-based query tuning through refinement/enlargement in a formal context. In J. Dix and S.J. Hegner (eds.), *Foundations of Information and Knowledge Systems. Proceedings of the 4th International Symposium (FoIKS 2006). Budapest (Hungary), 14-17 February 2006*. Lecture Notes in Computer Science 3861. Berlin-Heidelberg: Springer. 278-293

[2] Spyratos N. and Christophides V. 2005. Querying with preferences in a digital library. In K.P. Jantke, A. Lunzer, N. Spyratos and Y. Tanaka (eds.), *Federation over the Web. Proceedings of the International Workshop. Schloss Dagstuhl (Germany), 29 March – 1 April 2005*. Lecture Notes in Artificial Intelligence 3847. Berlin-Heidelberg: Springer. 130-142.

[3] Georgiadis P. 2005. Foundations of preference-based queries. In Proceedings of the 4th Hellenic Data Management Symposium (HDMS 2005). Athens (Greece), 25-26 August 2005.

[4] Koutrika G. and Ioannidis Y. 2004. Personalization of queries in database systems. In *Proceedings of the 20th International Conference on Data Engineering (ICDE 2004). Boston (USA), 30 March - 2 April 2004*. Washington: IEEE Computer Society. 597-608.

[5] Ioannidis Y. and Koutrika G. 2005. Personalized systems: models and methods from an IR and DB perspective. In K. Böhm, Ch.S. Jensen, L.M. Haas, M.L. Kersten, P.-Å. Larson and B.C. Ooi (eds.), *Proceedings of the 31st International Conference on Very Large Data Bases (VLDB 2005). Trondheim (Norway), 30 August - 2 September 2005*. New York: ACM Press. 1365.

[6] Koutrika G. and Y. Ioannidis 2005. Constrained optimalities in query personalization. In F. Ozcan (ed.), *Proceedings of the ACM SIGMOD International Conference on Management of Data. Baltimore (USA), 14-16 June 2005*. New York: ACM Press. 73-84.

[7] Koutrika G. and Y. Ioannidis 2005. Personalized queries under a generalized preference model. In *Proceedings of the 21st International Conference on Data Engineering (ICDE 2005). Tokyo (Japan), 5-8 April 2005*. Washington: IEEE Computer Society. 841-852.

[8] Koutrika G. and Y. Ioannidis 2005a. A unified user-profile framework for query disambiguation and personalization. In P. Brusilovsky, C. Callaway and A. Nürnberger (eds.), *Proceedings of the Workshop on New Technologies for Personalized Information Access (PIA 2005), part of the 10th International Conference on User Modeling (UM 2005). Edinburgh (UK), 24-29 July 2005*. 44-53.

Audio/Visual and Non-traditional Objects

Cluster objectives

Digital libraries will capture, organize, store and manage the access to large amounts of digital information regarding human knowledge, culture, and history in various, possibly interconnected, presentation forms like video, audio, images, etc. The overall long term research objective for this cluster is to formalize and develop a *Semantic Multimedia Management Framework* supporting the Audio-Visual aspects of the overall DELOS vision for Digital Libraries as well as the global demonstrator for future Digital Libraries being built by DELOS. Core research targets for this framework include: Metadata Capturing for Audio-Visual Content, Universal Efficient Access and Interactions with Audio-Visual Libraries, and effective Management of the Audio-visual Content. The framework identifies the following functionalities that are important for any future Audio Visual Library infrastructure:

Audiovisual Content Metadata Management: Functionality referring to the capturing of metadata for audiovisual content (semantic, structural, media, etc.). The metadata will be produced both manually and automatically. This can be achieved through the utilization of constructs provided by well accepted standards (such as MPEG 7/21) as well as domain knowledge captured in domain ontologies (such as OWL/RDF).

Audiovisual Content Access and Personalization: Functionality referring to advanced techniques for selection, delivery and personalization of the multimedia content, based on well-structured user preferences. In this context, transcoding and transmoding processes must be taken into account to cope with user requirements and resource availability. Another important objective here is the provision of multimodal query interfaces, search engines for efficient content retrieval.

Openness & interoperability support for domain-specific applications: Specification and formalization of the proper interface that will allow exploitation of audiovisual libraries by domain specific applications (e.g. e-Learning, cultural heritage, etc.)

Cluster activities

The most important activities foreseen in the workpackage include:

- Establishing common functionalities and advancing the state of the art in the area of metadata capturing from audiovisual content, including the investigation of issues related to multimodal information extraction, and the use of domain specific, context specific, and historical information in the extraction process.
- Establishing common foundations and advancing the state of the art in the area of information access and interactions with audio-visual digital libraries exploring multimedia content standards, domain and context specific knowledge, and investigating advanced interactions and interfaces to multimedia content.
- Establishing common foundations and advancing the state of the art in the area of management of audiovisual content, including new database models and data structures for storage, retrieval, and dissemination of multimedia data in emerging architectures and applications.

Cluster coordinators

Stavros Christodoulakis, Technical University of Crete, Greece, stavros@ced.tuc.gr

Alberto Del Bimbo, University of Florence, Italy, delbimbo@dsi.unifi.it

Video Annotation with Pictorially Enriched Ontologies

Marco Bertini, University of Florence, bertini@dsi.unifi.it
Stavros Christodoulakis, Technical University of Crete, stavros@ced.tuc.gr
Rita Cucchiara, University of Modena and Regio Emilia, rita.cucchiara@unimo.it
Alberto Del Bimbo, University of Florence, delbimbo@dsi.unifi.it
Costantino Grana, University of Modena and Regio Emilia, grana.costantino@unimore.it
Carlo Torniai, University of Florence, torniai@micc.unifi.it
Chrisa Tsinaraki, Technical University of Crete, chrisa@ced.tuc.gr

Keywords:.H.3.7 [**Information Storage and Retrieval**]: Digital Libraries; H.2.4 [**Systems**]:Multimedia databases

Task Members: D.S.I – MICC Univeristà degli Studi di Firenze, D.I.I Università degli Studi di Modena, Technical University of Crete, University of Amsterdam

Introduction

The ultimate goal of this Task is to automatically extract high-level knowledge from video data, permitting the automatic annotation of videos. In order to obtain effective annotation (both in the manual and automatic cases), one must rely on a domain-specific ontology defined by domain experts. The ontology is typically defined by means of a set of linguistic terms capable of describing high-level concepts and their relationships. However, often it is difficult to appropriately describe all interesting highlights in terms of (a set of) concepts only. Particularly in sport videos, while we can appropriately use concepts to describe basic types of highlights, like goal, counterattack, etc., it must be recognized that each one might occur in multiple contexts, each of which worthy of its own individual description. Distinguishing subclasses of these occurrences must be identified in order to group together instances that share the same or similar spatio-temporal characteristics.

Objectives of task

This task aims at defining methodologies and techniques to describe concepts and their specializations by augmenting an ontology of linguistic terms with "visual concepts" that represent these instances in a visual form. As an example consider the two soccer highlights shown in Figure 2 where an action of shot on goal follows two different patterns, and the description of the patterns using words would require long sequences that make impractical to retrieve the actions.

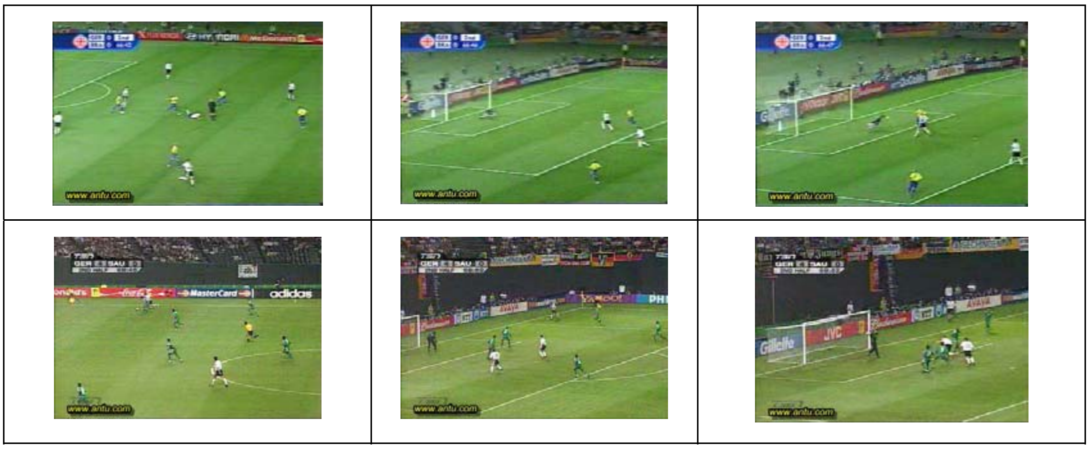


Figure 2 –in the first row is shown an attack action starting in the central-low part of playfield. Strong kick bounced back by the goal keeper and slow kick that scores a goal; in the second row is shown a fast attack action starting in the upper part of the playfield. Chip pass to the lower part of the field and header toward goal post that scores a goal

In this case it is more appropriate to use visual examples to represent a category of event patterns. Visual examples assume the role of visual concepts as an alternative to concepts expressed in linguistic terms.

Results

A set of tools has been developed within the task support dynamic creation and update of multimedia ontologies, provides facilities to automatically perform annotations and create extended text (and audio) commentaries of video sequences, and allows complex queries on video databases, based on the ontology itself.

Multimedia ontologies are expressed in the OWL standard, so that they can be shared and used in a search engine to perform content-based retrieval from video databases or to provide video summaries, and created using the GraphOnto tool developed by TUC/Music.

The linguistic part of the ontology is composed by a number of classes, that express the main concepts of the domain (e.g. actors, objects, facts and actions, highlights…) and their relationships. The extended multimedia ontology is created by linking video sequences as instances of concepts in the linguistic ontology, and performing an unsupervised clustering of the instance clips. An example of a multimedia ontology for the soccer domain is shown in Figure 2.

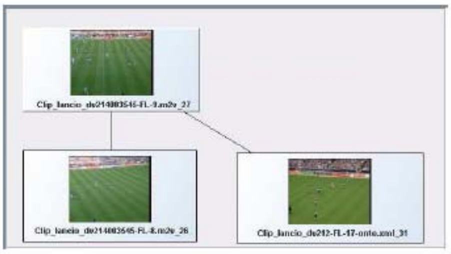
Annotation is performed at two distinct levels. At the clip level, the video sequence is segmented into clips, and each of them is annotated by checking its similarity with the visual concepts of the ontology. As the similarity with a particular visual concept is assessed, then higher level concepts linked to it in the ontology are immediately associated with the clip.

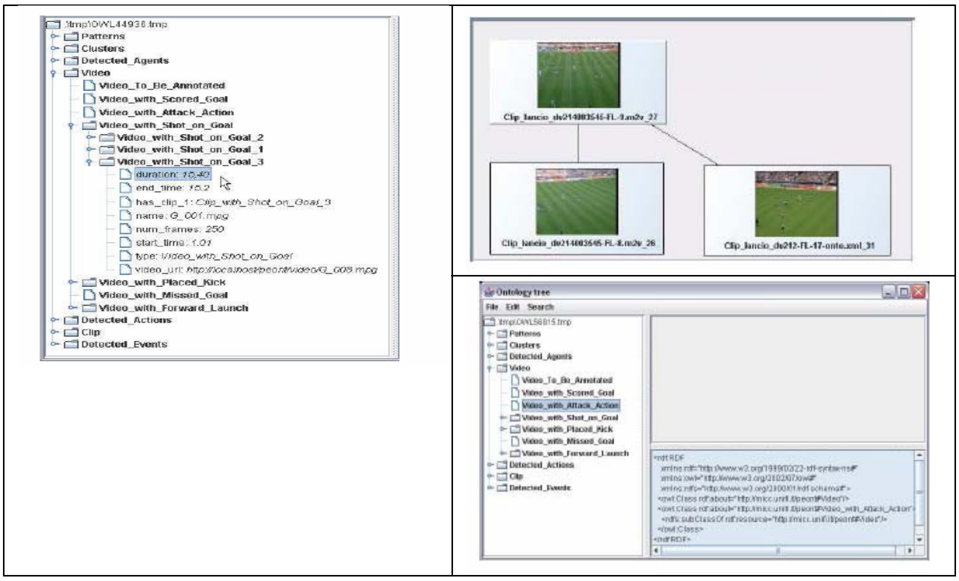
At the sequence level, composite concept patterns and the RACER description logic reasoner can be used to annotate a sequence of clips with some pre-defined articulated sentence. By reasoning on composite concept patterns, and using also the clip visual descriptors (which provide information about the playfield zone, the motion intensity of the action and the number of players involved), more extended commentaries can be created and associated automatically to video sequences.

The extended multimedia ontology is created by linking video sequences as instances of concepts in the linguistic ontology, and performing an unsupervised Fuzzy C-means clustering of the instance clips. Visual features that are used for clustering are both generic visual attributes (e.g. trajectories, motion fields, edge and color histograms computed from image data…) and domain specific descriptors (e.g. spatio-temporal feature combinations...) that qualify special events. The centers of the clusters are regarded as visual concepts, each representing a specific pattern in which a fact or event can manifest. A special class Undetected event/fact) is also created, that holds all the clips that are not classified as concept instances up to a pre-defined confidence.

The ontology includes concepts like crowd cheering, or referee action, such that their occurrences (e.g. crowd after a goal, crowd in a normal condition and crowd during attack action; referee yellow card, referee red card and referee fault signaling) can be easily distinguished each other, using generic features. Other more complex events like highlights, or player actions use instead more complex and domain-specific descriptors. The generic visual features used are: the color histogram (256 bins, in the HSV color space); the spatial color distribution (64 bins, obtained as the mean of colors in the YCbCr color space, for a 8x8 superimposed grid); the DCT coefficients (64 elements, according to the MPEG-7 specification for the Color Layout Descriptor); the average motion vectors (4 elements, obtained as the median of the MPEG motion vectors in each quarter of frame). The domain-specific descriptors used to model soccer highlights are: the playfield area; the number of players in the upper/lower part of the playfield; the camera motion intensity, direction and acceleration. For each clip, two feature vectors are created, respectively with six and four distinct components, one for each descriptor used, each component being a vector of as many elements as the number of frames in the clip, holding the values of the descriptor for each frame.

Composite concept patterns that are defined from temporal and semantic relations between the concepts in the ontology, and that are characteristic of the application domain, can also be included in the ontology to be used to ease annotation of long video sequences and express complex queries.

	 Clip_lancio_dv214003545-FL-9.m2v_27 Clip_lancio_dv214003545-FL-8.m2v_26 Clip_lancio_dv212-FL-17-onto.xml_31



```
.\tmp\OWL44938.tmp
  Patterns
  Clusters
  Detected_Agents
  Video
    Video_To_Be_Annotated
    Video_with_Scored_Goal
    Video_with_Attack_Action
    Video_with_Shot_on_Goal
      Video_with_Shot_on_Goal_2
      Video_with_Shot_on_Goal_1
      Video_with_Shot_on_Goal_3
        duration: 15.40
        end_time: 15.2
        has_clip_1: Clip_with_Shot_on_Goal_3
        name: G_001.mpg
        num_frames: 250
        start_time: 1.01
        type: Video_with_Shot_on_Goal
        video_url: http://localhost/peont/video/G_008.mpg
    Video_with_Placed_Kick
    Video_with_Missed_Goal
    Video_with_Forward_Launch
  Detected_Actions
  Clip
  Detected_Events
```

Clip_lancio_dv214003545-FL-9.m2v_27
Clip_lancio_dv214003545-FL-8.m2v_26
Clip_lancio_dv212-FL-17-onto.xml_31
Ontology tree
File
Edit
Search

```
.\tmp\OWL58815.tmp
  Patterns
  Clusters
  Detected_Agents
  Video
    Video_To_Be_Annotated
    Video_with_Scored_Goal
    Video_with_Attack_Action
    Video_with_Shot_on_Goal
    Video_with_Placed_Kick
    Video_with_Missed_Goal
    Video_with_Forward_Launch
  Detected_Actions
  Clip
  Detected_Events
```

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
 <owl:Class rdf:about="http://micc.unifi.it/peont#Video"/>
 <owl:Class rdf:about="http://micc.unifi.it/peont#Video_with_Attack_Action">
  <rdfs:subClassOf rdf:resource="http://micc.unifi.it/peont#Video"/>
 </owl:Class>
</rdf:RDF>
```

Figure 3 - the tree view of the ontology, the cluster pane and the OWL code pane

MIMA : Multimedia Interfaces for Mobile Applications

Margherita Antona, FORTH-ICS, antona@ics.forth.gr
Rita Cucchiara, University of Modena and Reggio Emilia, rita.cucchiara@unimo.it
Alberto Del Bimbo, University of Florence, delbimbo@dsi.unifi.it
Giuseppe Santucci, University of Rome “La Sapienza”, giuseppe.santucci@dis.uniroma1.it

Keywords:.H.3.7 [**Information Storage and Retrieval**]: Digital Libraries; H.2.4 [**Systems**]:Multimedia databases

Task Members: UNIFI-MICC, ROMA1, FORTH-ICS, UNIMORE

Introduction

The project investigates several strictly interrelated sub-problems, producing results in the framework of multimedia access for video presentation on mobile devices. The main subjects of investigation are:

- Automatic video extraction of meaningful objects and events according to user’s interests.
- User profiling and design of flexible small screen device interface, able to minimize the user interaction and adapt to devices characteristics.
- Performance measures and quantitative/qualitative indexes of user experience and satisfaction.

The overall project scenario is: the user is equipped with a PDA in order to receive multimedia information, i.e., video, images, graphic objects, text, and audio. The foreseen field of application is transmission of sports and news video, enriched with video summaries. The overall architecture is composed of three subsystems: Video annotation; Video summarization; User Interface.

Video annotation: Off-line annotation takes place on uncompressed video, producing a more precise annotation, extracting highlights and significant objects/events. Highlights must be represented with appropriate knowledge models based on the a-priori knowledge of the spatial-temporal structure of events and recognized by a model checking engine, based on statistical or model-based classification frameworks.

Construction of video summaries upon user’s request are managed by the Video summarization subsystem. Summaries are obtained dynamically, combining the user request with the annotations obtained from the off-line annotation process.

The User interface subsystem is in charge of handling the interaction with the user, and is faced with two main problems/objectives:

1. it should nicely fit the device characteristics and the user preferences.
2. it should include new interaction and visualization techniques to effectively convey the information produced by the annotation and video summarization systems.

Performance measure to assess the above aspects, the project explores suitable metrics able to measure the quality of different representations and choosing the solution that optimizes usability, user experience and required connection time.

Objectives of task

From the overall project scenario described before it appears that the role of the demonstrator is that of producing a new presentation from some raw video data. For example the selection of interesting highlights from a video, and the transcoding of the video, from a high bandwidth one to a lower bandwidth that can be transmitted on mobile networks, according to some user preferences.

For example let us a consider a sports writer working in a hotel is following the FIFA World Cup. In particular he is using his laptop PC to work on video sequence, and he needs to retrieve the shots that contain interesting highlights, browsing and searching them. Additionally we could add also a sub-scenario in which the journalist wants to access to the live camera in his apartment by using his mobile device. In general this part of the scenario fits perfectly with the main goals of the task delineated in the previous section. Only low bandwidth connections (GPRS- or UMTS-based) are available and display capabilities need specific interfaces for adapting the content to both the user’s preferences (in terms of viewing quality, color and spatial resolution, fluidity of the video and frame rate, etc.) and the device capabilities (display size, computational power, etc.).

To deliver videos in an ubiquitous manner, three key features are needed (see Fig. 1):

- automatic annotation of videos to provide tools for user’s selections of interesting entities and to summarize video content for content-based video adaptation;
- encoding/transcoding of video entities by streaming to fit in a very limited bandwidth (30-90 kbps);
- developing user interfaces suitable for mobile devices and user profile adaptation, e.g. by means of plastic interfaces.

Results

The data flow of the prototypal version of the demonstrator is shown in Fig. 1. Suppose the user wants to access to a sport video of a whole soccer game or a whole F1 race. Automatic annotation has been developed to extract semantics from both these types of videos. UNIFI-MICC has developed two systems for video annotation for soccer videos; a system is able to perform off-line annotation but results in higher performance in terms of precision and recall of detection of highlights such as shots on goal and penalty kicks, while the other works in almost real-time, detecting only a subset of the highlights. Since the second system works directly on MPEG videos, to achieve it speed, it was chosen to perform an initial integration with MPEG semantic transcoding system developed by UNIMORE, to test the feasibility of the system. Another integration effort, carried on within the MIMA project, between UNIFI-MICC and UNIMORE regarded the integration of annotation results obtained by the first system and a semantic video transcoder, that use results of the video annotation and compress the video according to content and user preferences.

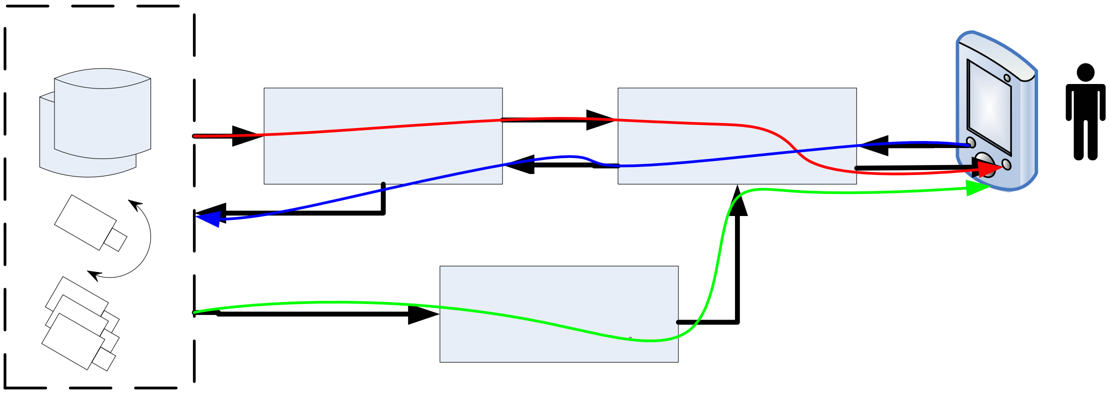


Figure 1 - Architecture of the demonstrator

The goal of obtaining automatic and human-supervised annotation of soccer videos is being carried on also within the DELOS VAPEON task (3.6) by UNIFI-MICC and UNIMORE, and further integration between the systems is foreseen.

UNIMORE created the PEANO system for extracting keyframes and shot/subshot boundaries from Formula 1 and other type of sport videos. The annotated video is processed by the user plastic interface under development by ROMA1 and FORTH-ICS] (red path (1)) to provide flexible device interfaces able to minimize the user interaction and improve data accessibility. FORTH-ICS has provided an internal deliverable that states the general UI heuristics and principles of “Universal Access” and “Design for all” to serve as a theoretical input for the design and development of the mobile’s prototype user interfaces. FORTH-ICS is also preparing a short report on methodologies and architectures to support appropriate UI adaptations. By means of the designed interface, the user specifies his selections in terms of classes of entities (objects and events) and corresponding weights (determining the relevance for the user). The user selections are fed back (blue path (2)) to the video annotation and summarization module to extract interesting video entities from the repository (in the case of stored videos) or to provide semantics for content-based video adaptation (in the case of live and stored videos). These video entities are processed by the video transcoding and streaming module (green path (3)) developed by UNIMORE in order to deploy them to the mobile device through the user plastic interface.

Concerning the plastic user interface, the ROMA1 prototype relies on a model describing the information that is exchanged between the user and the system foreseeing different ways of presenting the same information exchange on different physical devices.

The model consists of two main parts:

1. a set of Abstract Interaction Units (AIUs) to be used as building blocks for abstract interface definition and
2. the UML Activity Diagram as formalism to connect the AIUs that compose the interface. Using this approach the designer can specify the information content of each interaction and then connection between the various parts, designing the behaviour of the application.

The set of AIUs has been produced analyzing the user interfaces that are actually used to model standard web applications. Starting from specific interaction elements, we have grouped them into higher level units based on functional similarity. Such units express the key interactive features the specific elements of each group have in common.

Description, Matching and Retrieval by Content of 3D Objects

Stefano Berretti, University of Florence, berretti@dsi.unifi.it
Mohamed Chaouch, INRIA, mohamed.chaouch(a)inria.fr
Rita Cucchiara, University of Modena and Regio Emilia, rita.cucchiara@unimo.it
Alberto Del Bimbo, University of Florence, delbimbo@dsi.unifi.it
Mohamed Daoudi, University of Science and Technology of Lille, daoudi@enic.fr
Costantino Grana, University of Modena and Regio Emilia, grana.costantino@unimore.it
Pietro Pala, University of Florence, pala@dsi.unifi.it
Frank Seinstra, University of Amsterdam, fjseins@science.uva.nl
Jean-Philippe Vandeborre, University of Lille 1 – Sciences and Technologies, vandeborre@enic.fr
Anne Verroust, INRIA, anne.verroust@inria.fr
Marcel Worring, University of Amsterdam, worring@science.uva.nl

Keywords: H.3.7 [**Information Storage and Retrieval**]: Digital Libraries; H.2.4 [**Systems**]:Multimedia databases

Task members: Università degli Studi di Firenze (UNIFI-MICC), Italy - INRIA Roquencourt (INRIA), France - Università degli Studi di Modena e Reggio Emilia (UNIMORE), Italy - University of Amsterdam (UVA), The Netherlands - Univerité de Lille (USTL), France.

Introduction

Beside image and video databases, archives of 3D models have recently gained increasing attention for a number of reasons: advancements in 3D hardware and software technologies – in particular for acquisition, authoring and display – their ever increasing availability at affordable costs, and the establishment of open standards for 3D data interchange (e.g. VRML, X3D).

Thanks to the availability of technologies for their acquisition, 3D models are being employed in a wide range of application domains, including medicine, computer aided design and engineering, and cultural heritage. In this framework the development of techniques to enable retrieval by content of 3D models assumes an ever increasing relevance. This is particularly the case in the fields of cultural heritage and historical relics, where there is a growing interest in solutions enabling preservation of relevant artworks (e.g. vases, sculptures, and handicrafts) as well as cataloguing and retrieval by content. In these contexts, retrieval by content can be employed to detect commonalities between 3D objects (e.g. the “signature” of the artist), to monitor the temporal evolution of a defect (e.g., the amount of bending for wooden tables), to support services for tourists and visitors of historical sites (e.g., assist tourists in finding information related to an object of interest given a sample photograph of the object).

A major difficulty for the development of a system for retrieval by content of 3D objects relates to the need to capture the twofold nature by which 3D objects are experienced by humans: view based and structural. The ultimate goal of this task is to develop a system to support structural as well as view based retrieval by content of 3D objects. In this context, the project aims at the investigation of models for extraction of view based and structural based descriptors, models for indexing and similarity matching of structural and view based descriptors, models and metaphors for querying archives of 3D objects. The theoretical investigation of these models will end up with the design and development of a prototype system supporting structural and view based retrieval of 3D objects.

Work done

During the first 18 months of the task, research activities aimed to investigate existing solutions for 3D content based retrieval and propose innovative retrieval approaches. This is resulted in the definition of a set of new 3D object description and matching solutions capable to effectively support retrieval by content.

- 3D Objects representation

Three original solutions for representation and matching of 3D objects have been defined.

In [1] a view based approach has been defined exploiting the information that can be extracted from multiple views of a 3D object. In this method, an initial set of 2D views (projections) is created by considering equally spaced views on the unit sphere. To this end, models are first scaled so as to fit to a unit sphere centered in the barycenter of the 3D model. To select equally spaced positions for the views, a two units icosahedron centered in the origin is used. Zernike moments are then used to represent these views, and a clustering algorithm is employed to reduce the number of views. Finally, a probabilistic approach permits to match a query model against a models collection .

A view based approach is also proposed in [4] by using an enhanced depth buffer. In this case, models are first aligned using continuous PCA so as to fit into a unit cube. Then, projection images of the object on the six faces of the cube are used to characterize the object shape. These images are depth images whose pixels intensity represent the distance of the object from every pixels. A spectral analysis based on Fourier transformation is then used to describe and compare these depth images.

Following a different approach, in [2] the salient information of a 3D mesh is captured in the form of spin images signatures. A *spin image* at a mesh vertex is built mapping any other mesh vertex onto a two-dimensional space. Grey-level spin images are derived by considering the density of mesh vertices that map on the same point of the spin image, and evaluating the influence of each vertex over the neighboring pixels of its projection, according to a bilinear interpolation scheme. A compact descriptor is then extracted from the spin-images and used to reduce their number through a clustering algorithm.

In order to perform a comparative evaluation of these approaches, a set of reference methods for 3D objects retrieval have been implemented. In particular, the description techniques presently implemented include *3D geometric moments*, *curvature histograms*, *shape functions* and the MPEG-7 s*hape spectrum*. These have been combined with several distance functions used to evaluate the similarity between 3D objects descriptions during retrieval. Currently, these distance measures include the *Minkowski distance*, *Histogram Intersection, Kullbach-Leibler divergence* and χ^2 *distance.*

- Objects partitioning

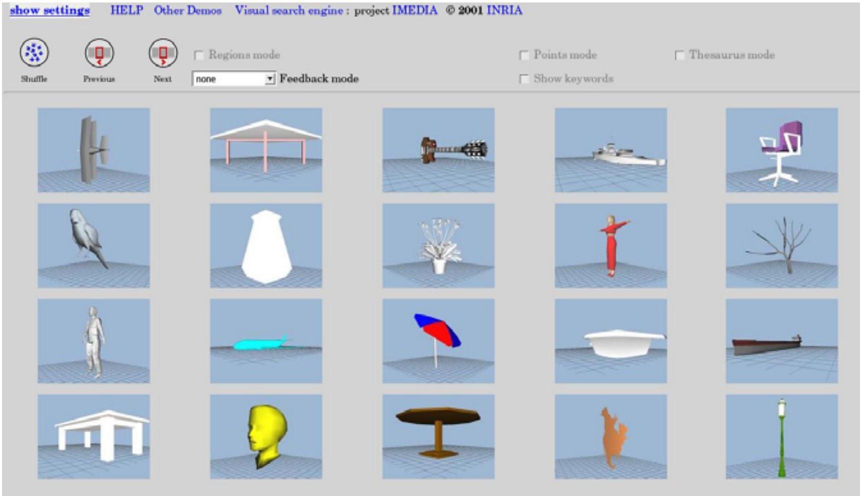
In the perspective to support part based retrieval of 3D objects, the automatic identification of salient object parts is a task of primary importance. To this end, different approaches for mesh partitioning (including *watershed, curvature clustering* and *oscillator networks*) have been experimented in order to decompose a 3D model into its constituent salient parts. This study has evidenced that topology based methods can provide effective objects partitioning. In fact, this kind of approaches tend to do not detect patches originated by slight curvature changes which can instead occur using curvature based methods. In particular, a segmentation approach based on the construction of a Reeb-graph and its successive reduction based on graph topological information has been developed and tested [3].

- 3D objects reconstruction

A further activity carried out under this task aims to develop capturing guidelines and methods allowing robust and complete reconstruction of 3D objects and scenes from video data obtained using a handheld camera. Initial software tools for each of the steps in 3D structure reconstruction, namely feature detection, matching, and building point cloud models have been developed, while methods for rectification and texture mapping are currently being developed. The result will be of general use for any digital library containing videos or collections of images showing the same object. 3D reconstruction can ultimately be used to create a collection of 3D models of large of collections of buildings or objects all over the globe, providing virtual visits to all of them.

- Task demonstrator

The work carried out by the task partners has been integrated in a prototype demonstrating the capabilities of a system supporting retrieval by content of 3D objects. The system features a Web accessible interface (publicly available at http://www-rocq.inria.fr/cgi-bin/imedia/3d.cgi) that allows users to browse a database of 3D objects. Browsing can be accomplished either by random sampling or by retrieval by visual similarity. Figure 4 shows the system interface with a set of objects randomly selected.



□ Regions mode

none Feedback mode

□ Points mode

□ Show keywords

□ Thesaurus mode

Figure 4: The user interface of the 3D content based retrieval demonstrator.

References

[1] Ansary T.F., Daoudi M. and Vandeborre, J.-P. forthcoming. A Bayesian 3D search engine using adaptive views clustering. In *IEEE Transactions on Multimedia*.

[2] Assfalg J., Bertini M., Del Bimbo A. and Pala P. forthcoming. Content based retrieval of 3D objects using spin image signatures. In *IEEE Transactions on Multimedia*.

[3] Berretti S., Del Bimbo A. and Pala P. 2006. Partitioning of 3D meshes using Reeb-Graph. In Proceedings of the International Conference on Pattern Recognition (ICPR 2006). Hong-Kong (China), 20-24 August 2006.

[4] Chaouch M. and Verroust-Blondet A. 2006. Enhanced 2D/3D approaches based on relevance index for 3D-shape retrieval. In *Proceedings of the International Conference on Shape Modeling and Applications. Matsushima (Japan), 14-16 June 2006*.

Automatic, Context-of-Capture Based, Categorization, Structure Detection and Segmentation of News Telecasts

Arne Jacobs, University of Bremen, jarne@tzi.de
George T. Ioannidis, University of Bremen, george.ioannidis@tzi.de
Martha Larson, Fraunhofer Institute for Media Communication, matha.larson@imk.fraunhofer.de
Stavros Christodoulakis, Technical University of Crete, stavros@ced.tuc.gr
Nektarios Moumoutzis, Technical University of Crete, nektar@ced.tuc.gr

Keywords: Information search and retrieval, modeling multimedia data, ontology design, speech recognition, structural models, video analysis

Research problem

Although video search technology is making rapid strides forward, it continues to be challenged by the semantic gap. This is the demanding problem of relating low-level features to the higher-level meaning that corresponds to the human-like understanding of video content. A solution to it is necessary for effective retrieval performance.

Our approach to bridging this semantic gap is twofold. First, we choose our application domain to be news videos, and second, we exploit the combination of multimodal analysis with semantic analysis based on ontologies. It is important to note that although we focus on news, the developed methods can be applied also to other genres. However, a significant part of video data produced today is news. News are broadcasted periodically, mostly every day, and only used once. There are many broadcasters that have developed and used their own news format, some broadcasters even focus completely on news.

A key observation in bridging the semantic gap in the news video domain is that semantic concepts in news videos are conventionalized in many ways. This fact can be exploited. For example, segments in news telecasts do not appear in arbitrary order, but rather follow a relatively strict scheme that determines the order of segments [1]. One often-used scheme is the separation of different news stories by anchor shots containing the presentation of the story that follows. These conventions, resulting in a distinct telecast structure, allow the viewer to easily recognize different segments. Each news format has its own structural model.

Proposed approach

We assume that the structural models underlying every news broadcast format can be described with context-free grammars[9]. Such a grammar can aid the segmentation process by removing ambiguities in classification, or by associating certain audiovisual cues with segment classes (eg news story, presentation). It models all interesting constraints in visual, audio and textual features, that are partly due to the way news programs are produced and partly to the habits and preferences of journalists and other agents related to the news production.

Following [2], we call the process of inferring the structure of a video broadcast ‘parsing’. For parsing of a news video following a corresponding model, we propose a system consisting of several recognizer modules and a parser modules. The recognizer modules analyze the telecast and each one identifies hypothesized instances of ‘events’ in the audiovisual input. Such events can be higher-level concepts like a specific person appearing in a shot (eg the anchor), the appearance of a certain series of frames (eg the introduction sequence, with which many news broadcasts commence), or low-level concepts, eg the similarity of two frames.

The system contains three distinct recognizer modules: the audio recognizer, the visual recognizer and the semantic recognizer. The visual recognizer identifies video events in the news stream, such as a face appearing at an expected position in the video, the occurrence of an anchor shot [3] or the presence of a familiar frame according to the expected structure of the broadcast. The audio recognizer identifies audio events such as the presence of speech [4] or music, the detection of predetermined keywords [5] and clustering of speakers. Finally, the semantic recognizer identifies the semantics involved in the telecast. This includes topic detection, high-level event detection, discourse cues [6] and possible story segmentation points.

The recognizers normally only communicate with the parser in a one-way communication, providing a sequence of predetermined feature ‘tokens’. The semantic recognizer is an exception, since that module requires a transcript of the telecast in order to perform its analysis. In the case where the transcript is not provided through the input (eg in the form of closed captions), the audio recognizer provides this information.

A *stochastic parser* [7] using a probabilistic grammar analyzes the identifications provided by the recognizers. In essence, the recognizers provide the parser with actual lexical tokens just as a lexical analyzer would provide to a programming language parser. The grammar represents the possible structures of the news telecast, so the parser can identify the exact structure of this telecast. When the parsing is complete and all the structural elements of the input have been analyzed, the semantic recognizer uses that information to identify story topics and events, and to assign all required semantics to the structure tree.

The grammar for each broadcast station, even for different news programs of the same station, is distinct. This is because the grammar captures the directing elements of the broadcast, and no two programs have exactly the same directional structure. Therefore, a grammar must be produced manually for each program examined.

To determine the probability values of the rules in the grammar, it is necessary to (currently manually) complete a training process, which uses a set of correctly labelled news recordings in the form of a sequence of tokens.

Finally, the semantic recognizer has access to an upper ontology, covering all necessary aspects required for multimedia content description, as well as domain-specific ontologies created for news. The ontologies are built on top of the OWL/MPEG-7 compatibility ontologies provided by [8]. The concepts acquired from these ontologies will define those detectable semantics that can be identified in the telecast.

Outlook and Current Demonstrator

The creation of a news format model is still a relatively time-consuming task, and currently has to be done mostly manually. It includes identification of the news format’s characteristic audiovisual cues (the ‘tokens’), the creation of a number of example token sequences, and the definition of the corresponding grammar capable of creating these sequences. Training of the probabilistic grammar can then be done automatically based on the example sequences.

To reduce manual efforts, we are currently investigating methods for automatically determining the audiovisual cues that characterize a given news format, by analysing a set of example recordings of that format. Based on the experience that these audiovisual cues do not change frequently, we expect videos from the same format to have many nearly identical video and audio sequences at similar time-points. We are planning to exploit this by using an inter-video, intra-format similarity analysis. We expect this analysis to result in two things: First, a number of example sequences will be returned for each identified token, which can be used to automatically train detectors for these token. Second, example token sequences will be generated, which will be usable for the generation and training of the probabilistic grammar.

The current demonstrator of the system consists of the probabilistic context-free grammar parser implemented as an extension of the popular JavaChart parser 0. The parser operates on simulated input containing the tokens of the recognizers. After the parsing of this simulated input using manually created grammar, the parse tree is provided that contains in its non-terminal symbols the information of the most probable segmentation of the news telecast along with a rough classification of the segments in genres (e.g. news stories, advertisements, weather forecasts, intro, outro, etc.). The final integration of the recognizers in the demonstrator will be based on the well-defined interfaces. In the final demonstrator, a full version of the semantic analyzer along with the fully developed News Ontologies and the semantic description creator module will be used to provide the final MPEG7 compiant XML description of the segmented and semantically analysed news telecasts.

References

[1] Bankert L., Jacobs A., Miene A., Hermes Th., Ioannidis G.T. and Herzog O. 2005. An environment for modelling telecast structures. In T. Catarci, S. Christodoulakis and A. Del Bimbo (eds.), *AVIVDiLib’05. Proceedings of the 7th International Workshop of EU Network of Excellence on Audio-Visual Content and Information Visualization in Digital Libraries. Cortona (Italy), 4-6 May 2005*. DELOS Network of Excellence. 176-179.

[2] Swanberg D., Shu C.-F. and Jain R. 1993. Knowledge guided parsing in video databases. In W. Nyblack (ed.), *Proceedings of the IST/SPIE Conference on Storage and Retrieval for Image and Video Databases*. IST/SPIE 1908. 13–24.

[3] Jacobs A. 2006. Using self-similarity matrices for structure mining on news video. In G. Antoniou, G. Potamias, C. Spyropoulos and D. Plexousakis (eds.), *Advances in Artificial Intelligence. Proceedings of the 4th Hellenic Conference on AI (SETN 2006). Heraklion (Greece), 18-20 May 2006*. Lecture Notes in Computer Science 3955. Berlin-Heidelberg: Springer. 87-94.

[4] Larson M. and Eickeler S. 2003. Using syllable-based indexing features and language models to improve German spoken document retrieval. In *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH 2003). Geneva (Switzerland), 1-4 September 2003*. 1217-1220.

[5] Osang, S. Entwicklung eines Schlüsselworterkennungssystems zur Medienbeobachtung. Diploma Thesis. Fachhochschule Bonn-Rhein-Sieg and Fraunhofer IMK.

[6] Maybury M.T. 1998. Discourse cues for broadcast news segmentation. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL 1998). Montreal (Canada), 10-14 August 1998. San Francisco: Morgan Kaufmann Publishers. Vol. 2: 819-822.

[7] Ivanov Y. and Bobick A.F. 2000. Recognition of visual activities and interactions by stochastic parsing. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(8): 852-872.

[8] Tsinaraki Ch., Polydoros P. and Christodoulakis S. 2004a. Integration of OWL ontologies in MPEG-7 and TVAnytime compliant Semantic Indexing. In A. Persson and J. Stirna (eds.), *Proceedings of the 16th International Conference Advanced Information Systems Engineering (CAiSE 2004). Riga (Latvia), 7-11 June 2004*. Lecture Notes in Computer Science 3084. Berlin-Heidelberg: Springer. 398-413.

[9] Stolcke, A. 1995. An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. In *Computational Lingistics* 21(2): 165-201.

[10] JavaChart Parser. The “NLP Farm” project. http://nlpfarm.sourceforge.net/javachart/

CoCoMA: Content and Context Aware Multimedia Content Retrieval, Delivery and Presentation

Elisa Bertino, University of Milan, bertino@dico.unimi.it
Susanne Boll, Susanne.boll@informatik.uni-oldenburg.de
Christian Breiteneder, Vienna University of Technology, breiteneder@ims.tuwien.ac.at
Stavros Christodoulakis, Technical University of Crete, stavros@ced.tuc.gr
Doris Divotkey, Vienna University of Technology, doris.divotkey@ims.tuwien.ac.at
Horst Eidenberger, Vienna University of Technology, hme@ims.tuwien.ac.at
Nektarios Gioldasis, Technical University of Crete , nektarios@ced.tuc.gr
Andrea Perego, University of Milan, perego@dico.unimi.it
Ansgar Scherp, Kuratorium OFFIS e.V., Ansgar.Scherp@OFFIS.DE
Chrisa Tsinaraki, Technical University of Crete, chrisa@ced.tuc.gr

Keywords: [H.2.4 Systems] Multimedia Databases, [H.5.1 Multimedia Information Systems] Video, [H.3.7 Digital Libraries] User issues, [H.3.3 Information Search and Retrieval] Information Filtering, Relevance Feedback, [H.5.1 Multimedia Information Systems] Video.

Research Problem

The increasing availability of high-speed wired and wireless networks as well as the development of a new generation of powerful (mobile) end-user devices like PDAs or cell phones leads to new ways of multimedia resource consumption. At the same time, new standards like MPEG-7/21 have become available, allowing us the enrichment of media content with semantic content annotations, which in turn facilitates new forms of multimedia experience, like search on specific topics or semantic-based content selection, filtering, and retrieval. *CoCoMA (Content and Context Aware Multimedia Content Retrieval, Delivery and Presentation)* focuses on the *integration* of *content and context-based multimedia retrieval from digital libraries* with the *personalized delivery and consumption* of the retrieved multimedia data.

Task Objectives and Followed Approach

The aim is to provide users of digital library systems with a solution for intelligent personalized retrieval from large media collections where media transport and presentation of the retrieval results are based on adaptation according to the user preferences. In particular he focus is on four major functions to be achieved:

- **Generic Presentation Authoring**, in order to support content-based personalized presentation of multimedia objects according to users' preferences and skill levels. In order to achieve this objective the MM4U framework [12], [4] is exploited. MM4U has been developed by OFFIS and is a generic and modular framework that supports multimedia content personalization applications.
- **Content-based Media Annotation and Retrieval**, for content-based metadata extraction and modeling, media annotation, query formulation and refinement, media access and user interface design. In order to achieve this objective the VizIR framework for content-based multimedia retrieval [5], [6], [7], [8], [9] is exploited. ViZIR has been developed by TUV and allows for content-based metadata extraction and modeling, media annotation (e.g. the entire MPEG-7 MDS), query formulation and refinement, media access and user interface design.
- **Semantics-based Media Annotation and Retrieval**, for describing and retrieving the multimedia content based on semantics. In order to achieve this objective, the DS-MIRF framework developed by TUC/MUSIC [10],[11],[13],[14],[15],[16],[17] is exploited. It allows for the interoperability of OWL with the complete MPEG-7 MDS so that domain ontologies described in OWL can be transparently integrated with MPEG-7 metadata.
- **Content Adaptation**, for adapting the multimedia content based on technical requirements coming by the execution environment (e.g. network bandwidth, end devices, etc.). For this purpose, the KoMMa [18] framework, developed by Klagenfurt University is utilized. It provides an open, extensible, and intelligent adaptation framework for multimedia data which can be used to build powerful multimedia adaptation servers or proxies. The framework makes use of MPEG-7 and MPEG-21 metadata and features a Prolog unit which is responsible for the adaptation decision taking process.
- **Multimedia Presentation Personalization**, for composing and delivering semantically rich personalized multimedia content to the users of an audiovisual library. In order to achieve this objective, the SyMPA multimedia authoring system (developed by UNIMI) [1],[2],[3] has been

exploited. SyMPA supports constraints for personalizing the presentation of multimedia objects according to users' preferences and skill levels. The execution flow is built dynamically on the basis of the semantic correlations existing among multimedia objects.

Current Demonstrator and Future Plans

The CoCoMA architecture utilizes and extends components (building blocks) provided by the frameworks accommodated in this task. These building blocks are all integrated in well-defined architecture which implements the CoCoMA Functionality. In a short, this functionality can be described as a high level process where multimedia metadata are produced (creation phase) and exploited for personalized multimedia consumption (consumption phase). Figure 5 sketches how this high level process is achieved in the currently developed demonstrator by illustrating the integration as well as the information/control flow among the various CoCoMA building blocks.

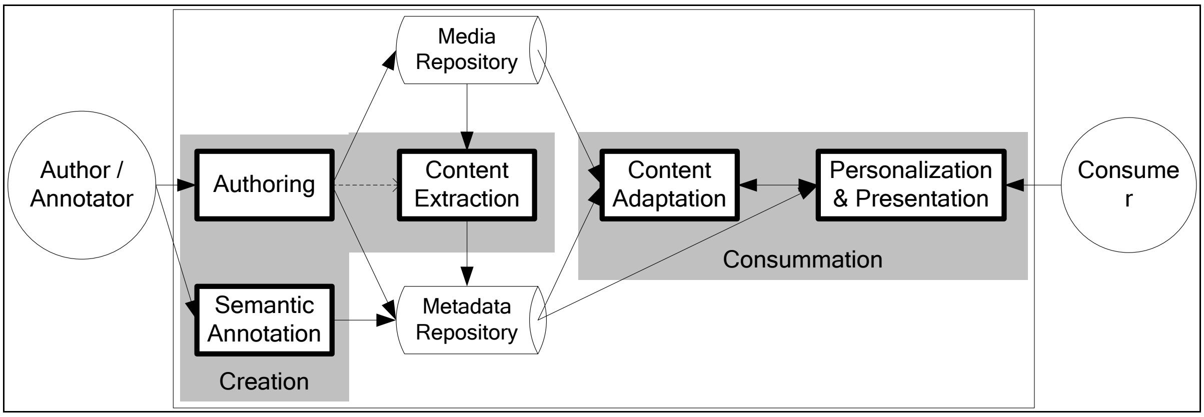


Figure 5: CoCoMA building blocks and workflow

The presentation authors and media annotators interact with the authoring building block and the semantic annotation interface. A knowledge base of ontologies supports the annotation process. The presentation consumers interact exclusively with the personalization and delivery component. Content-based annotation is controlled by the authoring process. Content adaptation is triggered by the presentation engine.

Media data and media metadata are organized in two major repositories. The media repository holds the temporal (e.g. audio, video) and non-temporal (e.g. text, image) media content specific to the presentation context. The metadata repository stores a variety of media-related metadata (semantic descriptors and low level media features in MPEG-7 XML) and user knowledge (e.g. semantic user preferences). Metadata includes non-temporal data (e.g. textual media segment descriptions, domain ontologies, presentation constraints) and temporal data (e.g. motion descriptions, spectral audio descriptions). The metadata repository is mostly fed by the two annotation building blocks and by the authoring process. Metadata is consumed by the content adaptation function (e.g. low-level color models, high-level relevance estimations) and by the personalization building block (e.g. merged with user knowledge for content-based media selection).

The next steps in CoCoMA Demonstrator include the integration of the implemented functionality into the Delos Digital Library Management System (DLMS). For this purpose, most its functionality is going to be provided (in JPA 3) following a service oriented approach. In particular, CoCoMA will be deployed as a set of well-defined and integrated (loosly coupled) web services which will be providing the current CoCoMA functionality to front-end tools and other Digital Library components through specific standard web service interfaces. A Description of the planned CoCoMA services can be found in [19].

References

[1] Bertino E., Ferrari E., Perego A. and Santi D. 2005. A methodology for the authoring of multi-topic multimedia presentations. In *AVIVDiLib'05. Proceedings of the 7th International Workshop of EU Network of Excellence on Audio-Visual Content and Information Visualization in Digital Libraries. Cortona (Italy), 4-6 May 2005*. DELOS Network of Excellence. 91-94.

[2] Bertino E., Ferrari E. and Stolf M. 2000. MPG: an interactive tool for the specification and generation of multimedia presentation. In *IEEE Transactions on Knowledge and Data Engineering* 12(1): 102-125.

[3] Bertino E., Ferrari E., Perego A. and Santi D. 2005. A constraint-based approach for the authoring of multi-topic multimedia presentations. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2005). Amsterdam (The Netherlands), 6-8 July 2005*. Washington: IEEE Computer Society. 578-581.

[4] Boll S. 2003. MM4U - A framework for creating personalised multimedia content. In Proceedings of the 9th International Conference on Distributed Multimedia Systems (DMS 2003). Miami (USA), 24-26 September 2003.

[5] Eidenberger H. and Breiteneder C. 2003. VizIR – A framework for visual information retrieval. In *Journal of Visual Languages and Computing* 14: 443-469.

[6] Eidenberger H. 2003. Media handling for visual information retrieval in VizIR. In T. Ebrahimi and Th. Sikora (eds.), *Visual Communications and Image Processing 2003***.** SPIE 5150. SPIE. 1078-1088.

[7] Eidenberger H. 2004. A video browsing application based on visual MPEG-7 descriptors and self-organising maps. In *International Journal of Fuzzy Systems* 6(3): 124-137.

[8] Eidenberger H. 2004. Statistical analysis of MPEG-7 image descriptions. In *ACM Multimedia Systems Journal* 10(2).

[10] Eidenberger H. and Divotkey R. 2004. A data management layer for visual information retrieval. In Proceedings of the 5th International Workshop on Multimedia Data Mining (MDM/KDD2004). Seattle (USA), 22 August 2004. 48-51.

[11] Kazasis F., Moumoutzis N., Pappas N., Karanastasi A. and Christodoulakis S. 2003. Designing ibiquitous personalized TV-Anytime services. In J. Eder, R. Mittermeir, B. Pernici (eds.), *Proceedings of the Ubiquitous Mobile Information and Collaboration Systems Workshop (UMICS 2003) held at the 15th Conference on Advanced Information Systems Engineering (CAiSE 2003). Klagenfurt and Velden (Austria), 16-20 June 2003*. CEUR Workshop Proceedings 75. CEUR-WS.org.

[11] Pappas N., Kazasis F., Moumoutzis N., Tsinaraki Ch. and Christodoulakis S. 2003. Personalized and ubiquitous information services for TV programs. In *Proceedings of the DELOS Workshop on Multimedia Contents in Digital Libraries. Chania (Greece), 2-3 June 2003*. ERCIM.

[12] Scherp A. and Boll S. 2005. MM4U: a framework for creating personalized multimedia content. In S. Nepal and U. Srinivasan (eds.), *Managing multimedia*

[13] Tsinaraki Ch., Fatourou E. and Christodoulakis S. 2003. An ontology-driven framework for the management of semantic metadata describing audiovisual information. In J. Eder and M. Missikoff (eds.), *Proceedings of the 15th International Conference on Advanced Information Systems Engineering (CAiSE 2003). Klagenfurt (Austria), 16-18 June 2003*. Lecture Notes in Computer Science 2681. Berlin-Heidelberg: Springer. 340-356.

[14] Tsinaraki Ch., Papadomanolakis S. and Christodoulakis S. 2001. Towards a two-layered video metadata model. In A. Min Tjoa and R. Wagner (eds.), *Proceedings of the 12th International Workshop on Database and Expert Systems Applications (DEXA 2001). Munich (Germany), 3-7 September 2001*. Washington: IEEE Computer Society. 937-941.

[15] Tsinaraki Ch., Polydoros P. and Christodoulakis S. 2004. Integration of OWL ontologies in MPEG-7 and TVAnytime compliant Semantic Indexing. In A. Persson and J. Stirna (eds.), *Proceedings of the 16th International Conference Advanced Information Systems Engineering (CAiSE 2004). Riga (Latvia), 7-11 June 2004*. Lecture Notes in Computer Science 3084. Berlin-Heidelberg: Springer. 398-413.

[16] Tsinaraki Ch., Polydoros P. and Christodoulakis S. 2004. Interoperability support for ontology-based video retrieval applications. In P. Enser, Y. Kompatsiaris, N.E. O'Connor, A.F. Smeaton and A.W.M. Smeulders (eds.), *Image and Video Retrieval. Proceedings of the 3rd International Conference (CIVR 2004). Dublin (Ireland), 21-23 July 2004*. Lecture Notes in Computer Science 3115. Berlin-Heidelberg: Springer. 582-591.

[17] Tsinaraki Ch., Polydoros P., Kazasis F. and Christodoulakis S. 2005. Ontology-based semantic indexing for MPEG-7 and TV-Anytime audiovisual content. In *Multimedia Tools and Application Journal* 26(3): 299-325.

[18] Leopold K., Jannach D., Hellwagner H. 2004. A knowledge and component based multimedia adaptation framework. In Proceedings of the IEEE 6th International Symposium on Multimedia Software Engineering (ISMSE 2004). Miami (USA), 13-15 December 2004. Washington: IEEE Computer Society. 10-17.

[19] JPA2 Delos Deliverable D3.0.1: "Report on Demonstrators of JPA2 Activities of WP3 (Audio/Visual and Non-traditional Objects).

Natural Language and Speech Interfaces to Multimedia Repositories

Konstantin Biatov, Fraunhofer Institute for Media Communication, Konstantin.Biatov@imk.fraunhofer.de
Tiziana Catarci, University of Rome ‘La Sapienza’, catarci @dis.uniroma1.it
Stavros Christodoulakis, Technical University of Crete, stavros@ced.tuc.gr
Anastasia Karanastasi, Technical University of Crete, allegra@ced.tuc.gr
Stephen Kimani, University of Rome ‘La Sapienza’, Stephen.Kimani@dis.uniroma1.it
Joachim Köhler, Fraunhofer Institute for Media Communication, joachim.koehler @imk.fraunhofer.de

Keywords: [1.2.1 Applications and Expert Systems] Natural language interfaces; [I.2.7 Natural Language Processing] Speech recognition and synthesis, [H.2.7 Database Administration] Data warehouse and repository, [H.5.2 User Interfaces] Evaluation/Methodology; [H.1.2 User/Machine Systems] Human factors

Research Problem

In the Digital Libraries of the future, knowledge management will be of major importance. However, traditional interfaces are particularly inflexible and difficult to use because of the highly structured and complex structures of knowledge. Natural Language Interfaces (NLI) and speech interfaces become quite attractive for such environments. This is especially true when the multimedia repositories are accessed using mobile devices. In addition to having high importance on their own, coupling the NLI user interface style with the traditional GUI style increases the accessibility of the overall system both from the point of view of the users with disabilities and from the point of view of users working in difficult contexts such as environments of scarce lighting, or when they carry out contemporarily another task such as driving a car.

Objectives and Followed Approach

The objective of this Task is to provide principles, methodologies and software for the automation of the construction of natural language and speech interfaces to multimedia repositories. These interfaces include capabilities for querying, filtering and ontology driven interaction formulation. We will also provide a specific application demonstrator of natural language and speech interfaces to multimedia repositories for soccer video games and we will evaluate the approach with human subjects.

The overall technical objective is to automate as much as possible the construction of natural language interfaces to knowledge bases. It has been shown that the overhead of developing natural language interfaces to information systems from scratch is a major obstacle for the deployment of such interfaces [3]. In the proposed approach, the storage structures of the metadata are not specified. The metadata could be stored in a multimedia repository or they could be stored in relational systems provided that the inference mechanisms that support the knowledge manipulation language have been built on top of them. The natural language system will also have to take into account in addition to the concept (domain) ontologies, word ontologies (like WordNet) and the interface between the two [4].

The task investigates a theoretical basis of the proposed approach that utilizes domain ontologies to a) find how a user request in natural language can be converted to a (structured and semantically enhanced) query to address the underlying multimedia repository using the user profile and context, and b) to allow a ranking of the results based on domain knowledge and user profile information. The proposed approach explored by TUC/MUSIC within DELOS has been published in [1], [2].

Current Demonstrator And Future Plans

The main goal that the current demonstrator aims to accomplish is to allow end users to submit speech or NL based information retrieval requests against a multimedia repository containing soccer video games and get back identified items. Speech recognition, NL processing, disambiguation, and information retrieval processes should be made transparently to the user. To this end, a multi-layered and component based approach has been followed in order to organise the entire functionality into well-defined and re-usable building blocks each one performing a specific function towards the desired objective.

The reference architecture of the Natural Language and Speech Interfaces Demonstrator comprises the following three layers:

UI Layer, which accommodates components that allow users to submit information retrieval requests and get back results of the Digital Libraries. Requests can either be spoken and recognized by a Speech Recognition Component [5], or typed in a Natural Language style through a specific Graphical User Interface (GUI). Results are shown in a standard result pane of the GUI component.

Natural Language Processing Layer (OntoNL), which is responsible for bridging the gap between Natural Language Interaction style and specific Knowledge Retrieval Languages supported by underlying Knowledge Repositories of Digital Libraries. This layer has been developed by TUC and it exploits domain knowledge (kept in ontologies) in order to disambiguate natural language user requests (typed or recognized from speech) and to generate appropriate queries according to the interface of the underlying knowledge repository.

Digital Library Layer, which represents the Digital Library against which user requests are submitted. In the current demonstrator this layer is materialized by a Semantic Search Service that has been developed by TUC and provides semantic retrieval and personalization services following the DS-MIRF Framework [6] which allows interoperability between OWL and the complete MPEG-7 MDS. This Semantic Search Service is offered by an XML Metadata Repository which also follows the DS-MIRF Framework.

The implementation of this architecture has been achieved with strong integration efforts not only within task 3.9 but also between this task and task 3.11 (Content and Context Aware Multimedia Content Retrieval, Delivery and Presentation). Although overcame, interoperability issues do not allow a configurable deployment of the implemented components. For that, our future plans include [6] the implementation of this architecture in Service Oriented Approach in mind supporting also its smooth integration into Delos DLMS.

On the other hand, it is highly desirable to investigate the generality of the approach as well as its performance in a different type of application. This way we will evaluate if the approach and the software produced is general enough to be used for fast construction of user interfaces in different applications that access knowledge-based digital libraries.

For the speech recognition component new directions involve the development of a robust front-end processing module to allow speech input in adverse conditions. This involves research on the adaptation of the acoustic model to recognize higher recognition rates including gender classification of input utterance and utterance clustering.

Finally, the research in the human-computer interaction aspects will finish a detailed evaluation of the first application and also pursue an evaluation of the second application, including heuristic as well as user based evaluation.

References

[1] Karanastasi A. and Christodoulakis S. 2005a. OntoNL: An Ontology-based Natural Language Interface Generator for Multimedia Repositories. In T. Catarci, S. Christodoulakis and A. Del Bimbo (eds.), *AVIVDiLib'05. Proceedings of the 7th International Workshop of EU Network of Excellence on Audio-Visual Content and Information Visualization in Digital Libraries. Cortona (Italy), 4-6 May 2005*. DELOS Network of Excellence. 206-215.

[2] Karanastasi A., Zotos A. and Christodoulakis S. 2006. User interactions with multimedia repositories using natural language interfaces: an architectural framework and its implementation. In *Proceedings of the 4th Special Workshop on Multimedia Semantics (WMS 2006). Chania (Greece), 19-21 June 2006*.

[3] Reithinger N., Alexandersson J., Becker T., Blocher A., Engel R., Löckelt M., Müller J., Pfleger N., Poller P., Streit M., Tschernomas V. 2003. Smartkom – Adaptive and flexible multimodal access to multiple applications. In *Proceedings of the 5th International Conference on Multimodal Interfaces (ICMI-PUI 2003). Vancouver (Canada), 5-7 November 2003*.

[4] Vargas-Vera M. and Motta, E. 2004. AQUA – Ontology-based question answering system. In R. Monroy, G. Arroyo-Figueroa, L.E. Sucar and J.H. Sossa Azuela (eds.), Advances in Artificial Intelligence. Proceedings of the 3rd Mexican International Conference on Artificial Intelligence (MICAI 2004). Mexico City (Mexico), 26-30 April 2004. Lecture Notes in Computer Science 2972. Berlin-Heidelberg: Springer. 468-477.

[5] Biatov K. and Larson M. 2005. Speaker clustering via bayesian information criterion using global similarity constraint. In Kokkinakis G., Fakotakis N. and Dermatas E. (eds.), *Proceedings of the 10th International Conference on Speech and Computer (SPECOM 2005). Patras (Greece), 17-19 October 2005*. Moscow: Moscow State Linguistics University.

[6] JPA2 Delos Deliverable D3.0.1: "Report on Demonstrators of JPA2 Activities of WP3 (Audio/Visual and Non-traditional Objects)

User-Interface and Visualization

Cluster objectives

The notion of a Digital Library is currently associated with technological and scientific efforts to build, maintain, and use large collections of electronic documents. However, it can also be regarded as a cornerstone in the construction of an information-enriched environment. Once this broader perspective is adopted, a variety of problems arise which will have to be solved in order to ensure the usability and accessibility of this environment to different users with varying needs and capabilities for both professional and recreational purposes. The ultimate goal of the User-Interface and Visualization cluster is to develop methodologies, techniques and tools to establish a theoretically motivated and empirically supported frame of reference for designers and researchers in the field of user interfaces and visualization techniques for digital libraries, so to enable future DL designers and developers to meet not only the technological, but also the user-oriented requirements in a balanced way. Specific objectives of the WP are:

- To elaborate a common understanding of the role and scope of user interface research in the digital library area.
- To develop a theoretical framework for digital library user interface design.
- To develop user-centered methodologies, techniques and tools to be exploited by DL designers and developers.

Cluster activities

User Requirement related Activities

- Systematic study of user requirements. The different perspectives on a digital library are being analysed to relate them to the requirements and technical implementation options that emerge from the ongoing development projects of the NoE partners.
- Analysis of user-related aspects in the development and usage of a DL system. The analysis will not focus only on the DL end user but will also take into account other DL stakeholders such as librarians, content providers and maintainers. The DL life cycle will be related to functional and non-functional requirements.
- Characterization of DL users. The characterization will take into account that the user interface accords accessibility for all categories of users, including users with special needs. In addition, this cluster will also explore how users can exploit a multi-modal DL user interface to meet their particular needs.

User Interface and Visualization Design Activities:

- Development of a taxonomy of relevant context models. A language specification is being investigated, which shall encompass the pertinent characteristics and requirements of context models that were identified during the development of the taxonomy.
- Development of a comprehensive model for relevance criteria. The consequence of taking the usage situation/context into account results in rethinking the basic assumptions underlying most contemporary approaches to information filtering and retrieval. This should lead to more realistic definitions of "relevance".
- Development of a theoretical framework from which user interface designers/developers can design DL user interfaces. The designer/developer gathers various resources provided by the theoretical framework (e.g. methodologies and tools) and designs a DL user interface (e.g. tailored for some particular application domain). Moreover, taking into account that future DL solutions will have to provide integrated customizable components that cover the appropriate functionality needed in a given context, the ultimate goal is to develop a design methodology and guidelines that, starting from a generic user interface, will allow to define tailored technical solutions that can be implemented in a given scenario, starting from the users needs.

Cluster coordinator

Tiziana Catarci, University of Rome ‘La Sapienza’, Italy, catarci@dis.uniroma1.it

Visualization in DL systems (Relevance feedback)

Daniel Keim, University of Konstanz, keim@informatik.uni-konstanz.de
Tobias Schreck, University of Konstanz, schreck@dbvis.inf.uni-konstanz.de

Problem Specification

The major challenge in content-based multimedia retrieval is the so-called semantic gap between machine computed similarity and human perception. Multimedia objects, e. g., images, time series, or 3D models, are represented in the system by high-dimensional feature vectors consisting of plain characteristics of the objects such as color histograms, wavelet coefficients, Fourier descriptors, and the like. Objects are represented as points in a high-dimensional vector space. In principle, similarity between objects can be easily expressed using a metric distance function. But the main problem is that users identify similar objects on a semantically higher level, e.g. a user might be looking for images portraying the same person.

For many multimedia objects, there exists a wealth of algorithms for extracting characteristic feature vectors, all of them having justification, but each of them being differently suited for retrieval of a given type of object as requested by a given user. So, we conjecture the *feature selection problem* has to be solved in order to provide effective searching facilities in Digital Libraries. The problem arises from the fact that there exist many different low-level features to use for answering a given query by a given user, but not all of the available feature vectors capture the semantically intended similarity notion which the user has in mind. This problem can be successfully addressed by integrating the user in the retrieval process, by means of relevance feedback. The basic idea in relevance feedback is to integrate the user in the retrieval process such that the system may learn or infer from the user his or her intended similarity notions, and appropriately select or combine the feature-based similarity metric available in the retrieval system. Visualization is a promising approach for effectively integrating the user in the relevance feedback loop, offering intuitive visual approaches to express relevance judgments, or visually determine the best suited FV spaces to use for his or her information need. This task will investigate novel methods of combining relevance feedback mechanisms with advanced visualization concepts, making the object space as well as the feature space visually accessible for the user. The expected results will enhance multimedia retrieval in digital libraries in two ways: (1) The user will have a better insight into the internal data representation and thus, will be able to provide a more appropriate query formulation for a given scenario; (2) The user will have more sophisticated tools to interact with query evaluation, so that the expected outcomes will be obtained in less feedback cycles.

Advanced Visualization in Digital Libraries

We propose to integrate the Self-Organizing Map (Kohonen Map; SOM) algorithm for support of scatter browsing, visual relevance feedback, and feature-level exploration.

Scatter Browsing and Digital Library Summarization

The SOM is highly suited to support scatter browsing in large DL over which the user needs to obtain an effective overview. The SOM can be used to acquire example objects to query for, and to assess the distribution of objects throughout a previously unknown DL. In our work we successfully engineered SOMs for a range of

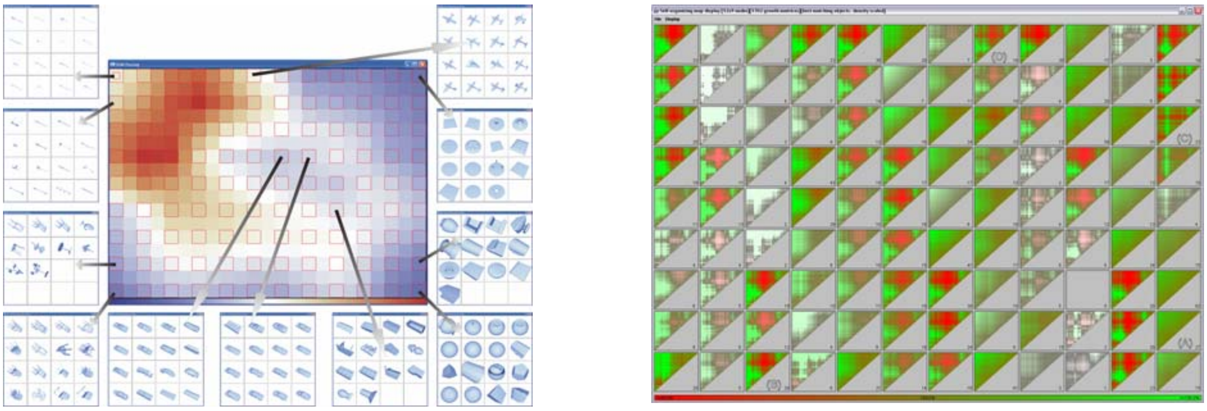


Figure 1: Application of the SOM algorithm on Digital Library content consisting of 3D objects (left) and time series data (right). The SOM is suited for object- and feature level visualization of DL content, and able to support feature selection, feature engineering, and relevance feedback for improved retrieval effectiveness.

different DL data types such as 3D models, E-Mail, and Time Series Databases. Figure 1 (left) illustrates a Self-organizing Map trained with the Konstanz 3D Benchmark, described using a discriminating feature vector. The map colouring reflects a density histogram, and the surrounding windows show database objects best-matching selected nodes on the map. Interestingly, the spatial distribution of elements over the map can be meaningfully interpreted in terms of gradual change in shape of the models. Figure 1 (right) gives a Self-organizing map of 1.700 growth matrices (a special visual representation of financial time series data). The map was learned from the down sampled original data. In both works, the input space is compressed in a meaningful way, allowing identifying salient global patterns in the data.

Visual Relevance Feedback

The SOM allows implementation of visual relevance feedback by integrating k-NN retrieval with free exploration of DL content organized on a SOM. Using appropriate visualization, mapping relevant objects to the SOM, the user is encouraged to explore interesting SOM regions, potentially revealing more relevant documents. We plan to design and evaluate different schemes of SOM-based relevance feedback. Reproducible, objective precision-recall experiments should be performed, allowing to contrast the SOM-approach with classical relevance feedback schemes from machine learning.

Unsupervised Discrimination Power Estimation

We found that the SOM offers also a wealth of valuable information regarding low-level feature spaces. We research heuristics allowing to assess the discrimination power to be expected for a DL represented in a given feature space. Preliminary results indicate this to be a promising direction for boosting the retrieval power in Digital Libraries content-based retrieval. Adapting existing global selection heuristics to work with user supplied relevance information, performing query-dependent combination of feature vectors is a challenging future work in this context.

Extensibility to Distributed Environments

Relevance feedback techniques have been usually studied and applied in centralized scenarios, where users and databases are located at an individual, local site. Actually, the increasing availability of multiple repositories accessible in a distributed environment over the Network, poses additional issues to relevance feedback. In particular, searching information through the Internet often requires users to separately contact several digital libraries, use each library interface to author the query, analyze retrieval results and merge them with results returned by other libraries. Such a solution could be simplified by using a centralized server that acts as a gateway between the user and several distributed repositories: The centralized server receives the user query, forwards the user query to federated repositories - possibly translating the query in the specific format required by each repository - and fuses retrieved documents for presentation to the user. To accomplish these tasks efficiently, the centralized server should perform some major operations, such as: *resource selection*, *query transformation* and *data fusion*. Resource selection is required to forward the user query only to the repositories that are candidate to contain relevant documents. In this operation, relevance feedback could be used to adaptively change the selection process. Also in the data fusion phase, relevance feedback could be used to modify the fusion process according to different relevances associated to the repositories in successive queries. In this scenario, we will investigate if current relevance feedback techniques, targeted for centralized applications, can be easily and effectively expanded in order to account for distributed contexts or if specifically tailored relevance feedback techniques should be developed.

Partners

National and Capodistrian University of Athens, Greece (UOA)

Università degli Studi di Firence, Italy (UNIFI-MICC)

Università di Roma “La Sapienza”, Italy (Roma1)

Institut National de Recherche en Informatique et en Automatique, France (INRIA)

References

[1] Bustos B., Keim D., Saupe D., Schreck T. and D. Vranic 2006. An experimental effectiveness comparison of methods for 3D similarity search. In *International Journal on Digital Libraries* 6(1): 39–54, 2006.

[2] Bustos B., Keim D., Saupe D., Schreck T. and D. Vranic 2005. Feature-based similarity object databases. In *ACM Computing Surveys* 37: 345–387.

T. Kohonen 2001. *Self-Organizing Maps. 3rd edition*. Springer Series in Information Sciences 30. Berlin-Heidelberg: Springer.

[3] Schreck T., Keim D. and Panse. C. Visual feature space analysis for unsupervised effectiveness estimation and feature engineering. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2006). Toronto (Canada), 9-12 July 2006*.

[4] Schreck T. forthcoming. *Effective Retrieval and Visual Analysis in Multimedia Databases*. PhD thesis, University of Konstanz.

[5] Smeulders A.W.M., Worring M., Santini S., Gupta A. and Jain, R. 2000. Content-based image retrieval at the end of the early years. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(12): 1349-1380.

User Requirements-driven Support for a DL Design Framework

Davide Bolchini, University of Lugano, davide.bolchini@lu.unisi.ch
Tiziana Catarci, University of Rome ‘La Sapienza’, catarci @dis.uniroma1.it
Stavros Christodoulakis, Technical University of Crete, stavros@ced.tuc.gr
Norbert Fuhr, University of Duisburg, fuhr@uni-duisburg.de
Mathieu Le Brun, Virtual Resource Centre for Knowledge about Europe − CVCE
Annelise Mark Pejtersen, CSE, ampcse@mail.dk

Keywords: H.4.3 [Information Systems Applications]: Communications Applications – Internet; H.5.4 [Information Interfaces and Presentation]: Hypertext/Hypermedia – user issues.

Research Problem

Despite the rapid ongoing technological evolution in recent years (the success of the World Wide Web, the diffusion of various kinds of interactive applications, and the availability of different end-user devices), DL interfaces are still based mainly upon “search” and “search refinement” mechanisms. During the first 18 months of the DELOS NoE, the work of WP4 “User Interfaces and Visualization” has focussed on identifying functional and non-functional requirements of Digital Libraries (DLs), with the aim to establish an empirical basis for user interface design for DLs and to define a taxonomy of functional DL infrastructure and visualization paradigms. Additionally, a preliminary DL usage lifecycle model has been elaborated, targeted to facilitate, in the light of the different usage phases that characterise the long-term life and evolution of DLs, further analysis of user requirements. The results of such an empirical study calls for the investigation of the potential effectiveness and benefits to the user stemming from a full adoption of alternative interaction paradigms, and especially of novel techniques for navigation such as browsing by catalogues, semantic linking, information visualization, interactive maps, social navigation, etc., which are seldom and occasionally employed in current DLs.

Objectives

This proposal aims at extending and enhancing the results of previous efforts in DELOS towards the systematic investigation of non-conventional interaction paradigms, and the correlation of such paradigms with different usage phases of DLs.

In particular, the research objectives to be pursued are the following:

1. Further extend the empirical analysis of user functional and non-functional requirements, and further analyse, refine and enhance the preliminary life-cycle model in the context of the continued collection and analysis of empirical data.
2. Specify the user requirements regarding novel interaction paradigms in addition to the ones traditionally employed, such as “search”, “querying” and their variations.
3. Define advanced interaction paradigms for DL and thereby build a “theoretical” framework for the design of new DL interfaces.
4. Develop a prototype (on top of already existing DL’s, or for new ones) demonstrating the new concepts and mechanisms.
5. Test the effectiveness and usability of the prototype against the needs of selected user communities.

Ongoing Research Activities

- Continuous investigation of existing literature on non-conventional access paradigms
- Refinement of Catalogue-Browsing paradigm
- Definition of user requirements (by means of scenario-based analysis) suitable for the paradigm
- Development of prototype demonstrating the paradigm at work
- Refinement and advancement of the research about requirements for “aural” access paradigms, particularly suitable for granting “accessibility” to information-rich DLs to visually-impaired users
- Definition of user requirements (by means of scenario-based analysis) suitable for an “accessible” paradigm to information
- Development of a prototype application demonstrating the paradigm at work (“page reader”).

- Refinement of an empirically motivated model of the user experience lifecyle in DLs (FORTH-ICS)
- Refinement of the Scatter-Browsing and of the corresponding prototyping applications

References

[1] Bolchini D., Colazzo S., Paolini P. forthcoming. Requirements for aural web sites. In *Proceedings of the 8th IEEE International Symposium on Web Site Evolution (WSE 2006). Philadelphia (USA), 23-24 September 2006*.

[2] Bolchini D. and Paolini P. 2006. Interactive dialogue model: a design technique for multi-channel applications. In *IEEE Transactions on Multimedia* 8(3): 529-541.

[3] Bolchini D. and Paolini P. 2004. Goal-driven requirements analysis for hypermedia-intensive web applications. In *Requirements Engineering Journal* 9(2): 85-103.

[4] Bertini E., Catarci T., Di Bello L. and Kimani S. 2005. Visualization in digital libraries. In M. Hemmje, C. Niederee and T. Risse (eds.), *From Integrated Publication and Information Systems to Information and Knowledge Environments. Essays Dedicated to Erich J. Neuhold on the Occasion of His 65th Birthday*. Lecture Notes in Computer Science. Berlin-Heidelberg: Springer. 183-196.

[5] Mirabella V., Kimani S., Gabrielli S. and Catarci T. 2004. Accessible e-Learning material: a no-frills avenue for didactical experts. In *New Review of Hypermedia and Multimedia* 10(2): 165-180.

[6] Antona M., Mourouzis A., Kartakis G. and Stephanidis. C. 2005. User requirements and usage life-cycle for digital libraries. In G. Salvendy (ed.), *Human Computer International 2005. Proceedings of the 11th International Conference on Human-Computer Interaction (HCI International 2005). Las Vegas (USA), 22-27 July 2005*. Mahwah: Lawrence Erlbaum Associates.

[7] Tombros A., Malik S. and Larsen B. 2005. Report on the INEX 2004 interactive track. In *ACM SIGIR Forum* 39(1): 43-49.

[8] Tombros A., Larsen B. and Malik S. 2005. The interactive track at INEX 2004. In N. Fuhr, M. Lalmas, S. Malik and Z. Szlávik (eds.), *Advances in XML Information Retrieval. Revised Selected Papers from the 3rd International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2004). Dagstuhl Castle (Germany), 6-8 December 2004*. Berlin-Heidelberg: Springer. 410-423.

Task-centered Information Management

Tiziana Catarci, University of Rome 'La Sapienza', catarci @dis.uniroma1.it
Alan Dix, University of Lancaster, alan@hcibook.com
Benjamin Habegger, University of Science and Technology of Lille, habegger@grappa.univ-lille3.fr
Yannis Ioannidis, National and Kapodistrian University of Athens, yannis@di.uoa.gr
Akrivi Katifori, National and Kapodistrian University of Athens
Anton George Lepouras, National and Kapodistrian University of Athens
Antonella Poggi, University of Rome 'La Sapienza', poggi@dis.uniroma1.it

Research problem and objectives

Most existing human-computer interfaces are now based on the WIMP (Windows, Icons, Menus, Pointers) paradigm and on the desktop metaphor. Each application is targeted towards editing one specific type of "document". Current desktop oriented systems propose a mostly disconnected set of generic tools (word processor, e-mail reader, video and image visualizer, etc.). Let us consider the following Scenario. John Smith is a 35 year-old sports writer working with SILC (Sport International Library Center). He is currently staying in Ibis Hamburg Wandsbek hotel near the FIFA World Cup Stadium in Hamburg, where he has been following and covering the World Cup championships. He normally receives news updates from the SILC. In particular, as soon as a new match finishes, he is notified with an email message, having as object the name of the competing nations as well as the match result. Each time he receives such a notification email, he therefore opens the Word Processor and starts writing an email to SILC. To do so, he looks for videos and images about the match that just ended, he visualizes selected videos and images, and possibly he annotates those that mostly interest him. Finally, when John finishes the article, he attaches the article to the email and send it to SILC. Therefore, in order to achieve the task of writing an article commenting the last World Cup championships match, he has to interact with many different, autonomous tools. The fact that these tools are running on one system without being connected leads to awkward situations such as John reentering or copying the name of the competing nations between the different applications. Furthermore, John has to run this same task, in a rather similar manner, each time he receives the notification email from SILC. Having a system which is aware that some task is being run would allow for some anticipation and could relieve the user from much routine procedures.

This is precisely the goal of Task 4.8, which aims at providing the user with a module having knowledge of his specific tasks. The idea is that such a task-oriented module should be designed to aid the user in managing his *day-to-day tasks* and helping him gain in efficiency. We consider that both the tasks themselves and how they are to be lead are specific to each user. Building a module which is both adapted to the user and simple enough for a non-expert to use is a challenge. At one extent, full adaptability is asking the user to program himself his tasks, while at the other extent, simplicity is having preprogrammed tasks not necessarily adapted to the user's needs. Therefore, having some form of adaptability, can not go without at least some participation of the user in defining his tasks. However, combining user interaction and inference methods (in particular machine learning) can help in building a system making it easier for the user to provide task descriptions.

This work follows up our previous work on the OntoPIM system [5], which has concentrated on the management of the user's data through a so-called Personal Ontology. The latter provides the user with a virtual description of his domain of interest, through which the user can access his personal real data [2]. Thus, ontology-based querying offers a useful basis to define user tasks, in that the Personal Ontology can be used to specify how to map task input from and/or output data to the DL (e.g. by posing an appropriate query over the Personal Ontology). In other terms, we propose a task definition language allowing to possibly specify the task data flows in terms of queries over the Personal Ontology, i.e. over the DL.

Results

As already mentioned, in the last period, the goal of this task has been to design a system mainly focused on user's tasks, by extending the OntoPIM framework to make it handle such tasks. One of the main target objectives of the system is to improve user efficiency by minimizing user input and maximizing automation. Much importance has been given to having a user-centered system and in particular one which could be used by non-expert users. This has lead to consider using different inference mechanisms allowing to (partially) detect user tasks. The main results of this work are given below. For more details, see [1].

First, we have given a task definition language that is characterized by a well-defined semantics. Our proposal is based on the idea of combining task decomposition and a plan language to describe for each task, the execution plan of its subtasks. Here we only give an intuition of our specification language. On one hand, a task is decomposed into a set of subtasks. This allows for a more comprehensible view of the task. For example, John may have defined his “Comment World Cup Match” as depicted in figure 2.1. In this example, the task has been decomposed into three higher grain subtasks : “Write Artical”, “Manage Match Information”, and “Send Article”. These reflect the way in which John works when he is writing a sport article : he writes the article by simultaneously retrieving and inspecting images from different sources about the event, and finally he sends it to SILC. Each of these subtasks can again themselves be divided. On the other hand, we propose a plan language limited to sequence, alternatives and repetition. Our language is somewhat related to workflow languages used to model business processes such as YAWL. The restrictions keep the language simple enough for inference to remain feasible and allows to have a quite straight forward semantics. Also, from the user's point of view, a useful benefit of such languages is that they have a natural graphical representation which is often more comprehensible for a non-expert user.

Second, for the task specification process to both be simple for the novice user and provide some power to the more expert user, we have proposed two different approaches to build a task specification by a combination of user interaction and techniques based on machine learning. The first approach allows the user to himself define a task decomposition. Then, given example runs of each subtask, the system is responsible for inferring missing parts of the overall task specification. On the other hand, in the second approach task-aware environment keeps track of user actions and identifies frequent patterns. Then, from these patterns new task specifications are proposed to the user, to gradually build up task hierarchies.

Finally, we have specified how to extend the OntoPIM architecture in order to manage tasks. Basically, to do so, we envisage to include a monitoring subsystem, an execution subsystem, an inference engine, a task repository and a presentation subsystem. Note that the monitoring subsystem is crucial for the system, in that it is responsible for tracking user events and keeping a precise trace of the user’s actions. Indeed, task specifications are derived starting from the log produced by the monitoring system.

Future work and integration

Many directions can be envisaged for the future. First of all, we envisage to build a prototype of a TIM-centered system combining a task definition module and a logging module. This would allow to test and compare the different inference mechanisms we have proposed. Furthermore, by making the system available to users, it would allow having feedback on the ideas developed.

Also, in the current approach, tasks are specified using a simple hierarchy of subtasks and actions. Having an ontology of tasks might be beneficial in defining and managing a growing number of tasks. For example, such an ontology could allow to categorize the users tasks and/or allow to define new tasks on the basis of existing related tasks. Finally, having a task oriented architecture rises many questions on how the system should interact with the user. Using currently existing tools as basic actions, will likely lead to awkward situations for the user such, such as giving input in separate dialogs which could have been merged into one. Task-driven information management therefore rises the problem of designing interfaces in the context of TIM.

Task members

University of Roma 1 “La Sapienza”: Tiziana Catarci, Benjamin Habegger, Antonella Poggi

University of Lancaster: Alan Dix

University of Athens: Yannis Ioannidis, Akrivi Katifori, Georgios Lepouras

References

[1] Catarci T., Habegger B., Poggi A., Dix A., Ioannidis Y., Katifori V. and Lepouras G. 2006. Intelligent user task-oriented systems. In *Proceedings of the Personal Information Management SIGIR Workshop. Seattle (USA), 10-11 August 2006*.

[2] Catarci T., Dong L., Halevy A. and Poggi A. forthcoming. *Personal Information Management,* chapter *Structure Everything.* Seattle: University of Washington Press.

[3] Dix A., Catarci T., Habegger B., Ioannidis Y., Kamaruddin A., Katifori A., Lepouras G., Poggi A. and Ramduny-Ellis D. 2006. Intelligent context-sensitive interactions on desktop and the web. In *Proceedings of the Context in Advanced Interfaces Workshop (at AVI 2006). Venezia (Italy), 23-26 May 2006*.

[4] Dix A., Levialdi S. and Malizia A. 2006. Semantic halo for collaboration tagging systems. In S. Weibelzahl and A. Cristea (eds.), Workshop on Social Navigation and Community-Based Adaptation Technologies (held at the 4th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems - AH2006). Lecture Notes in Learning and Teaching. Dublin: National College of Ireland. pp. 514-521.

[5] Katifori V., Poggi, A. Scannapieco M., Catarci T. and Ioannidis Y. 2005. OntoPIM: how to rely on a personal ontology for personal information management. In S. Decker, J. Park, D. Quan and L. Sauermann (eds.), Proceedings of the ISWC 2005 Workshop on the Semantic Desktop - Next Generation Information Management & Collaboration Infrastructure. Galway (Ireland), 6 November 2005. CEUR-WS 175.

DiLAS: a Distributed Service for the Management of Annotations

Maristella Agosti, University of Padua, agosti@@dei.unipd.it
Hanne Albrechtsen, Institute of Knowledge Sharing, hanne.albrechtsen@knowshare.dk
Nicola Ferro, University of Padua, ferro@@dei.unipd.it
Ingo Frommholz, University of Duisburg-Essen, ingo.frommholz@uni-due.de
Preben Hansen, Swedish Institute of Computer Science, preben@sics.se
Emanuele Panizzi, University of Rome “La Sapienza”, panizzi@di.uniroma1.it
Annelise Mark Pejtersen, Center of Cognitive Systems Engineering, ampcse@mail.dk
Ulrich Thiel , Fraunhofer IPSI, thiel@ipsi.fraunhofer.de

This work reports on the status of advancement and development of the Digital Library Annotation Service (DiLAS) project which is aimed at designing and developing an architecture and a framework able to support and evaluate a decentralized annotation service. Specific attention is here given on the integration of the annotation service into DAFFODIL and BRICKS.

The mission of DiLAS is to foster change in users’ interaction with Digital Library Management Systems (DLMS) and contribute to developing services for social infrastructure in Digital Libraries (DL). This mission is addressed through the provision of a new independent annotation service for interactive knowledge creation and sharing. The annotation service enriches the DL contents and usage, the users personalize the information in a new contextual learning opportunity, and they collaborate by sharing this new knowledge.

All the project participants have previous experience in developing a number of annotation systems targeted to different application domain users. Among those systems, there are (in alphabetical order): BRICKS (Building Resources for Integrated Cultural Knowledge Services), COLLATE (Collaboratory for Annotation Indexing and Retrieval of Digitized Historical Archive Material), DAFFODIL (Distributed Agents for User-Friendly Access of Digital Libraries), FAST (Flexible Annotation Service Tool), IPSA (Imaginum Patavinae Scientiae Archivum: Image Archive of the Paduan School), and MADCOW (Multimedia Annotation of Digital Content Over the Web).

Building on those previous experiences the project participants have identified the use cases the DiLAS annotation service needs to support and have decided to design and develop a generic annotation service, that is a service to be used into different DLMS. To this end, the architecture of the DiLAS system has been defined and it consists of three layers – the data, application and interface logic layers. This decomposition allows us to achieve a better modularity within DiLAS and to properly describe the behaviour of DiLAS by means of isolating specific functionalities at the proper layer.

The data logic layer manages the actual storage of the annotations and provides a persistence layer for storing the objects which represent the annotation and which are used by the upper layers of the architecture. In order to make it as flexible as possible, an abstract API for the functionalities of the storage has been defined. This API, in turn, allows for accessing different systems to perform the actual storage of the annotations. In the first prototype of the DiLAS system we use the MADCOW system as actual storage for the annotations, but for the final prototype we are going to integrate also the BRICKS system as storage provider.

The application logic layer provides advanced functionalities that make use of annotations, such as for example the search and retrieval of annotations described above. As in the case of the data logic layer, we defined a set of abstract API that make the access to the DiLAS service functionalities independent from the particular implementation provided.

The interface logic layer is devoted to manage the interaction with the end-user. It depends on the system into which DiLAS is going to be used and relies on the DiLAS Abstract Service API in order to provide the functionalities described above to the end user.

For the first prototype of DiLAS we use the DAFFODIL system in order to carry out some of the described user-level use cases, but in the final prototype the BRICKS system will be supported as well. Note that these two systems offer different kinds of functionalities based on annotations, but these functionalities are obtained as composition of the functionalities provided by the DiLAS Abstract Service API.

Note that both the application logic (Abstract Service API) and the upper part of the data logic (Abstract Storage API) correspond to the respective layers in the FAST system. Thus, FAST represents the architectural framework which makes it possible to integrate MADCOW, DAFFODIL, or BRICKS together. Indeed, FAST describes both these layers and the business objects exchanged among these layers by means of abstract interfaces. Those interfaces provide us with a general framework for describing the interaction and integration of the different layers without coupling it with a specific implementation. As a consequence, the integration of MADCOW, DAFFODIL, or BRICKS requires to provide a concrete subclass of the FAST layers and business objects, in order to fit them to the needs of the newly integrated systems.

The DiLAS annotation model derives from the FAST model. The current implementation of the DiLAS model provides a partial set of functionalities in order to fit into the MADCOW model.

With respect to the chosen architecture, a first prototype demonstrator of the DiLAS system has been developed at the "interface logic" layer. The objective of this demonstrator is to allow an easy access to the DiLAS functionalities, while showing the log of all the system activities, because this demonstrator is intended to be used by system developers for testing purposes more than by end users.

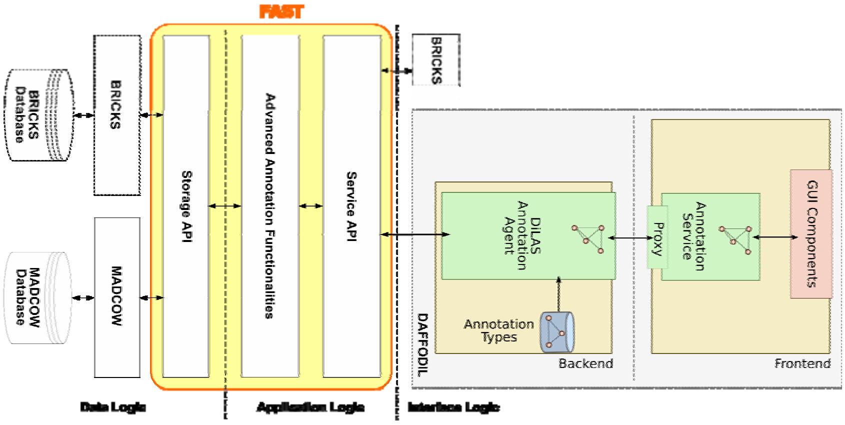


Figure 1: DiLAS/DAFFODIL architecture.

Figure 1 shows the architecture of DAFFODIL/DiLAS. DAFFODIL consists of two main parts: the backend running on a dedicated server, and the front-end executed on the user machine. A DiLAS annotation agent was implemented and is running in the DAFFODIL backend. This agent communicates with the FAST server.

On the DAFFODIL front-end, another annotation service is running which communicates with the DAFFODIL backend through a proxy. When the user creates a new annotation with the GUI components, the annotation service requests the annotation types in the ontology from the backend and creates a memory model of it. This memory model is used to dynamically set up the list of possible annotation types in the user interface. The newly created annotation, an instance of one of the annotation type classes, is sent to to backend annotation agent. The OWL code representing the annotation and its metadata is converted into a valid MADCOW format and sent to the FAST server. Similar for the other direction – if an annotation thread is requested for an object, the backend annotation agent queries the FAST server for all necessary annotations. These are transformed into instances of their corresponding annotation type class and the resulting OWL code (classes and instances) is sent via the proxy to the front-end annotation service in order to be displayed.

References

[1] Agosti M., Albrechtsen H., Ferro N., Frommholz I., Hansen P., Orio N., Panizzi E., Pejtersen A.M. and Thiel U. 2005. DiLAS: a digital library annotation service. In J.-F. Boujut (ed.), *Proceedings of Annotation for Collaboration - A Workshop on Annotation Models, Tools and Practices (IWAC 2005). Paris (France), 23-24 November 2005*. Paris: CNRS - Programme Société de l'Information. 91-101.

[2] Agosti M., Ferro N., Albrechtsen H., Frommholz I., Panizzi E. and Thiel U. 2005. Design, implementation and evaluation of the use of digital multimedia annotations in DL interaction and DL users collaboration. In C. Thanos (ed.), *DELOS Research Activities 2005*. DELOS Network of Excellence. 47-48.

[3] Agosti M., Ferro N., Frommholz I., Panizzi E., Putz W. and Thiel U. 2006. Integration of the DiLAS Annotation Service into Digital Library Infrastructures. In A. Blandford and J. Gow (eds.), *Proceedings of the. 1st International Workshop on Digital Libraries in the Context of Users' Broader Activities (DL-CUBA 2006). Chapel Hill (USA), 11-15 June 2006. 1-4*.

[4] Agosti M., Ferro N., Frommholz I. and Thiel U. 2004. Annotations in digital libraries and collaboratories – Facets, models and usage. In R. Heery and L. Lyon (eds.), *Research and Advanced Technology for Digital Libraries. Proceedings of the 8th European Conference (ECDL 2004). Bath (UK), 12-17 September 2004*. Lecture Notes in Computer Science 3232. Berlin-Berlin-Heidelberg: Springer. 244–255.

[5] Agosti M., Ferro N., Panizzi E. and Trinchese R. 2006. Annotation as a support to user interaction for content enhancement in digital libraries. In A. Celentano and P. Mussio (eds.), *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI 2006). Venice (Italy), 23-26 May 2006*. New York: ACM Press. 151-154.

Knowledge Extraction and Semantic Interoperability

Cluster objectives

The thematic area of Semantic Interoperability is growing in importance in digital library (DL) research (taking the interpretation of "digital library" at its broadest). It applies to the application of different vocabularies and terminology used in descriptions of digital objects for both learning and research, collections of those objects, collections of datasets and resources used in the wider cultural heritage sector and in e-research. Indeed, cross-sectoral and cross-domain shared understanding of semantic descriptions is one of the goals of the Semantic Web as envisaged by Tim Berners-Lee in his roadmap published in 1998. This vision has more recently (2001) been applied to Grid computing and e-science/e-research initiatives in the Semantic Grid approach. In addition, the application of algorithms for the mining and analysis of digital resources (text, data, complex objects), offers exciting opportunities for the extraction of new knowledge and the re-use of data and information in new ways. Today, we are beginning to address some of the issues and challenges in this complex area and the cluster has the opportunity to carry out some important research to move the Semantic Web/Grid vision forwards towards implementation. The two strategic goals are:

- To co-ordinate a programme of activities which brings together research excellence from a range of inter-related knowledge engineering and information management areas, and which facilitates the sharing of experience and expertise amongst practitioners from both DL and Grid/computing science backgrounds.
- To explore the potential of new models, algorithms, methodologies and processes in a variety of technical applications, institutional frameworks and cross-sectoral environments, which will lead to the creation of guidelines and recommendations of best practice for dissemination to the widest possible community of interest.

Cluster activities

A Forum has been created to provide a physical and virtual arena for the exchange of experience and research in all the areas/themes of this cluster. It provides an opportunity to integrate systematically other relevant groups into the cluster. This development is being supported by a moderated virtual forum or discussion list for the expansion of discussion on selected topics. It is also intended to maximise opportunities to harmonise with other relevant initiatives such as CIDOC and FRBR. In the near term future the cluster will concentrate on the following goals:

(i) To build links between the digital repository work and the digital library reference model developments.

(ii) To work more closely with e-Learning and cultural heritage user communities in the design, testing, evaluation and enhancement of the prototype systems and to facilitate their integration with the Delos Digital Library.

(iii) To extend the functionality of the *GraphOnto* tool and promote its application and use in other Delos research tasks as an exemplar of integration within the Network.

(iv) To investigate and develop methods and a demonstrator for the integration of heterogeneous data types, models, upper level ontologies and domain specific KOS .

Cluster coordinator

Elizabeth Lyon, UKOLN, University of Bath, U.K., E.Lyon@ukoln.ac.uk

Interoperability of eLearning Applications with Digital Libraries

Antonia Arahova, Ionian University, tonia@ionio.gr
Polyxeni Arapi, Technical University of Crete, xenia@ced.tuc.gr
Stavros Christodoulakis, Technical University of Crete, stavros@ced.tuc.gr
Sarantos Kapidakis, Ionian University, sarantos@ionio.gr
Haroula Konsolaki, Ionian University, hkonsolaki@yahoo.com
Nektarios Moumoutzis, Technical University of Crete, nektar@ced.tuc.gr
Manolis Mylonakis, Technical University of Crete, manolis@ced.tuc.gr
Christos Papatheodorou, Ionian University, papatheodor@ionio.gr
Manjula Patel, University of Bath, m.patel@ukoln.ac.uk
Barbara Vagiati, Ionian University, barbaravag@yahoo.gr

Keywords: digital libraries, interoperability, standards, distance learning

Research Problem

The most important application of Digital Libraries (DL) is to support knowledge and learning purposes. However, DLs and their standards have been developed independently of eLearning applications and their standards. For that, interoperability issues between digital libraries and eLearning applications are risen (complex and multilevel problem). In order to enable the construction of eLearning applications that easily exploit DL contents it is crucial to bridge the interoperability gap between digital libraries and eLearning applications. Task 5.4 is exploring the interoperability of eLearning applications and Digital Libraries looking particularly at data models, standards and workflows. The aim is to study the major standards for digital libraries (e.g. METS), eLearning (e.g. SCORM) and audio-visual content description (e.g. MPEG-7), and to produce mappings among them. Based on this, the objective is to develop an integration framework and a service-oriented architecture which will be validated by a specific demonstrator.

Task Objectives and Followed Approach

The task has sought to answer the following questions: (a) What are the major architectural requirements and workflows for effectively supporting eLearning applications running over digital libraries? (b) What are the major interoperability requirements for DL and eLearning standards? (c) What are the management requirements and tools for audiovisual material and 3D object representations, which form the basis for many collections of learning resources?

The task focused on the design and implementation of appropriate tools which can be deployed across the wider DL practitioner community. Initial work has addressed developing models for an architectural framework and workflow, producing mappings and transformations between relevant metadata standards, implementing the GraphOnto tool (Tsinaraki, Polydoros & Christodoulakis, 2005), implementing aspects of the architecture and documenting the issues in a series of reports.

In associated work, models for supporting semantic 3D information to be used in a variety of eScience applications have been derived and functionality requirements investigated. Two ontologies for 3D scenes, based on formal and de facto standards have been developed (Kalogerakis, Christodoulakis & Moumoutzis, 2006). They are available on DELOS WP3 testbeds and demonstrators site for downloading: http://astral.ced.tuc.gr/delos/. This work is complementary to the work in the Task 3.8 “Description, matching and retrieval by content of 3D objects”, which does not consider semantic descriptions.

Current Demonstrator and Future Plans

The Architecture for Supporting Interoperability between Digital Libraries and ELearning Applications (ASIDE) (Arapi, Moumoutzis & Christodoulakis, 2006) has been designed and implemented during JPA2. ASIDE addresses the identified interoperability problems in a layered architecture where eLearning (and other) applications are built on top of digital libraries and utilize their content. ASIDE offers a generic framework for the automatic creation of personalized learning experiences using reusable A/V learning objects. It is service-oriented and conforms to the IMS Digital Repositories Interoperability (IMS DRI) Specification. The IMS DRI specification provides recommendations for the interoperation of the most common repository functions enabling diverse components to communicate with one another: search/expose, submit/store, gather/expose and request/deliver. It is recommended that these functions should be implementable across services to enable them to present a common interface.

The ASIDE architecture consists of the following layers:

The **Digital Library Layer**, where digital objects are described using METS+LOM (eLearning context), and MPEG7 (A/V descriptions) building this way interoperable A/V learning objects, which can be transformed to SCORM and delivered to eLearning applications (METS/SCORM transformation component).

Applications Layer, where e-Learning Applications (e.g. Learning Content Management Systems, Learning Management Systems etc.) discover, access, and use the content of the A/V content of the digital library through appropriate services (resource utilizers). The generated personalized A/V learning experiences are delivered to the applications in the form of SCORM packages. Any SCORM-compliant system can recognize and "play" these packages.

The **Middleware Layer**, which consists of the following components:

a) The **METS/SCORM transformation component**, which is responsible for the transformation of the METS descriptions pointing to LOM and MPEG7 descriptions to SCORM Content Packages. This includes not only simple transformation from METS XML file to SCORM manifest file, but also the construction of the whole SCORM package (PIF). More-over, the mime-type of the files is taken into account and, if needed, intermediate html pages are constructed with links to these files (e.g. in case of video files).

b) The **Personalized Learning Experiences Assembler (PALEA)**, which, taking into account the knowledge provided by the Learning Designs (abstract training scenarios) and the Learner Profiles described later, constructs the personalized learning experiences and delivers them in the form SCORM Packages. The dashed arrow in the left side of PALEA indicates that using this component is optional and that digital library services can be directly accessed (e.g. a teacher wants to find appropriate learning objects to construct manually a learning experience).

Ontologies providing knowledge to the PALEA for the automatic construction of personalized learning experiences:

a) **Domain Ontologies** that provide vocabularies about concepts within a domain and their relationships.

b) **Instructional Ontology** that provides a model for the construction of abstract training scenarios. These are pedagogical approaches (instructional strategies/didactical templates), which can be applied to the construction of learning experiences. This ontology developed in T5.4 has the important characteristic that learning objects are not bound in the training scenarios on design time, as in current eLearning standards and specifications (e.g. IMS Learning Design and SCORM). Whereas, pedagogy is separated and independent from content achieving this way reusability of learning designs or parts of them that can be used from the systems for the construction of "real" personalized learning experiences, where appropriate learning objects according to the learner profile are bound to the learning experience at run-time.

Learning Designs are abstract training scenarios in a certain domain built according to the model given in the instructional ontology.

The **Learner Profiles** constructed using the vocabulary given in the Learner Profile Ontology, which represents a learner model for the creation of learner profiles. Elements from IEEE PAPI and IMS LIP specifications have been also used in this model. Some important elements of this model are: learner goals, competencies, previous knowledge, educational level and learning style.

The interoperability architecture has been implemented using the following technologies: Web services, JavaTM 2 Platform, Standard Edition, v1.5, Berkeley DB XML, Jena API, SPARQL RDF Query Language (Prud'hommeaux & Seaborne, 2005) and XQuery for querying the XML-based metadata descriptions of the digital objects stored in the digital library.

During JPA3, T5.4 plans to work on three major objectives: 1) implement further extensions to the GraphOnto tool, which are needed for metadata management on top of integrated repositories of various kinds (audiovisual, 3D graphics etc.). This objective directly contributes to the development of the SMMF envisioned in WP3 by providing an enhanced core component for the management of ontologies for audiovisual and non-traditional objects, 2) development of an Earth Sciences' digital library according to the T5.4 eLearning environment for the provision of eLearning in Geography, and 3) support of interoperability and semantics for 3D objects. This objective is complementary to Delos Task 3.8 and is related with the envisioned SMMF of WP3 by studying the semantic aspects of 3D objects and their integration in multimedia DLs.

References

[1] Arahova A. and Kapidakis S. 2005. Empowering our libraries, empowering our education system: using the research results for implementing not the best, but the most effective policy for school libraries. In Proceedings of International Federation of Libraries Associations Conference (IFLA 2005). Oslo (Norway), 14-18 August 2005.

[2] Arapi P., Moumoutzis N. and Christodoulakis S. 2003. Supporting interoperability in an existing e-learning platform using SCORM. In *Proceedings of the 3rd IEEE International Conference on Advanced Learning Technologies (ICALT 2003). Athens (Greece), 9-11 July 2003*. Washington: IEEE Computer Society. 388.

[3] Christodoulakis S., Arapi P., Moumoutzis N., Patel M., Kapidakis S., Arahova A. and Bountouri L. 2005. Interoperability of eLearning applications with audiovisual digital libraries. DELOS poster session in conjunction with ECDL 2005. Vienna (Austria), 18-23 September 2005.

[4] Kalogerakis V., Christodoulakis S. and Moumoutzis N. 2006. Coupling ontologies with graphics content for knowledge driven visualization. In *Proceedings of the IEEE Virtual Reality International Conference (VR 2006). Alexandria (USA), 25-29 March 2006*. Washington: IEEE Computer Society. 43-50.

[5] Tsinaraki C., Polydoros P. and Christodoulakis S. 2005. GraphOnto: a component and a user Interface for the definition and use of ontologies in multimedia information systems. In T. Catarci, S. Christodoulakis and A. Del Bimbo (eds.), *AVIVDiLib'05. Proceedings of the 7th International Workshop of EU Network of Excellence on Audio-Visual Content and Information Visualization in Digital Libraries. Cortona (Italy), 4-6 May 2005*. DELOS Network of Excellence. 99-102.

[6] JPA2 Delos Deliverable D5.4.2: "Demonstrator of mapping between the eLearning and AV content description standards.

Ontology-driven interoperability

Martin Doerr, FORTH-ICS, martin@ics.forth.gr
Douglas Tudhope, University of Glamorgan, dstudhope@glam.ac.uk

Rationale

Traditional Libraries provide access to documents via general subjects and metadata about the creator and creation of the document. The search paradigm is restricted to the retrieval of like documents with respect to some search criteria. Although advances have been made in search engine technology, Information Retrieval techniques and standard metadata schemes such as Dublin Core, current methods for integrating material from different domains and terminology systems remain poor, compared to the research demands and the results of manual human investigation.

The challenge for the next generation of information access systems is the ability to retrieve complementary objects and deep paths of relevant relationships that cross multiple document and resource boundaries. Intellectual, logical and physical architectures must be found to identify dynamically relevant resources for complementary information, and to reliably link multiple resources with mechanisms to disambiguate referred items, such as persons, places, objects, periods, but also types and other scientific concepts. Resources of general background knowledge such as gazetteers, VIP lists and domain ontologies should be accessible as automated information services providing necessary bits of information to close gaps in queries and information chains, such as placename-to-coordinate translations. This pertains to most domains in science, culture and business.

Complementary information can only be identified against an application and domain overarching core-ontology that allows for relating, mediating or translating the elements of the necessarily heterogeneous data and metadata schemata employed in multiple applications and domains. The promoters of the CIDOC CRM core-ontology (ISO21127) could prove that such ontologies can be created, that they can be kept extraordinarily generic supporting a kind of general discourse (macroscopic, discrete historical or retrospective analysis in this case), and that they can be fairly compact. It appears feasible to organize in the sequence the internal and external logical structure of wide research networks of DL of various disciplines and KOS services in a way that will allow for seamless access to relevant data paths across multiple resources.

Objectives

The activity addresses the key aim of achieving semantic interoperability at both data and metadata levels. Knowledge Organization Systems (KOS), such as classifications, gazetteers and thesauri provide a controlled vocabulary and model the underlying semantic structure of a domain for purposes of retrieval. Ontologies provide a higher level conceptualisation with more formal definition of roles and semantic relationships. The objective of this project is the investigation and development of methods for the integration of heterogeneous data types, models, upper level ontologies and domain specific KOS . This effort will be driven by a domain overarching core ontology starting from the CIDOC CRM (ISO 21127) and will be realised via research reports, guidelines, real world case studies and a demonstrator. Tasks selected for investigation will span the spectrum of applied to general focus. The experimental material will be taken from the particularly rich cultural heritage domain and traditional library science.

Description of Work

In more detail, the work comprises three aspects:

A) A collaboration between CIDOC CRM-SIG and IFLA-FRBR Review Group on the creation of a core ontology merging the FRBR and CIDOC CRM concepts. The harmonization of FRBR represents a widening of the scope of the CIDOC CRM from its origins in the Museum domain to Archives, Libraries and Museums on an international level. It has been carried out by a series of meetings of an interdisciplinary group of experts, and elaboration of results in between meetings. Resulting from this effort, a first draft of the FRBRoo model has been published http://cidoc.ics.forth.gr/docs/frbr_oo/frbroo_0-6-5.doc. Presentations on this effort were made by Martin Doerr and, on the FRBR side, by Patrick Le Boeuf (Bibliotec Nationale, Paris) and Maja Zumer (University of Ljubljana) at a workshop in March, on Semantic Interoperability for e-Research in the Sciences, Arts and Humanities at the Internet Institute, Imperial College, London - http://cidoc.ics.forth.gr/london_workshop.htm (ppts of cited presentations are available).

B) Semantic schema mapping prepares the ground for semantically rich and precise infrastructures to multi-purpose clusters of digital libraries. Whereas most semantic Web activities try to recover knowledge from bad and idiosyncratic structures, this work investigates systematic ways to produce the necessary diversity of ergonomically optimized data structures without losing meaning with respect to a common ontology. Work demonstrates two directions: the derivation of rich data structures for documentation units from the core ontologies on one side, and the derivation of minimal metadata structures for simplified querying on the other side. A mapping of DC Types to the CRM has been created. This is a significant step towards creating DC access to CIDOC CRM compatible resources, increasing semantic richness while maintaining the simplicity of DC.

C) In virtually all data record we find categories of things, frequently called “types”. Depending on the degree of detail a schema captures, some of those categories may be more analyzed than in others. E.g.: “type: clay pot” in one schema may appear in another as “material: clay, form: pot”. In order to overcome this problem a mapping from domain ontologies (or thesauri) to an overarching common schema, ie the CRM, is needed. It was decided to investigate the environmental archaeological domain as a test case, building on an English Heritage (EH) existing extension of the CRM to their needs. This specialisation of the CRM schema had only existed previously on paper. Working in Protégé and in collaboration with EH, Glamorgan augmented the standard CIDOC CRM 3.4.9 release with the environmental archaeology section (in fact, the majority) of the EH extended model. This can now be exported in various formats. EH have done a preliminary (satisfactory) assessment of the extended model. A report has been produced for EH, so that they are enabled to continue the work in their own right. Additionally, mappings have been made to the new EH Environmental Archaeology Thesaurus (EAT) and the (Archaeology Data Service’s) Environmental Archaeology Bibliography (EAB) database. In collaboration, the University of Lund has implemented a pilot demonstrator showing that mappings can enable search of the EAB (employing additional mappings to uncontrolled database fields). A presentation on the work was also made at the Semantic Interoperability Workshop mentioned – see (A) above. A research grant has recently been awarded to Glamorgan, by the UK AHRC Research Council, for a 3 year project exploring this general area.

Participants:

FORTH (leader), NTNU, DSTC, MTA SZTAKI DSD, Imperial College London, Ionian University Greece, Economic University of ATHENS, University of Glamorgan, University of Lund, Technical University Chania, IFLA FRBR Review Group, CIDOC CRM Special Interest Group.

Digital Preservation

Cluster objectives

This cluster focuses on enhancing preservation methods, tools and functions within the context of digital libraries. It has broad goals to promote the adoption of preservation technologies in digital library development designs, to raise the profile of digital preservation issues within the Digital Library Community, and to increase collaboration with other international researchers conducting research within the digital libraries and preservation communities. Among the core activities in this area there is a summer school on Digital Preservation in Digital Libraries; the last edition was held in San Miniato in June 2006; a summary report about it is in press with the journal *Archivi e Computer* and a fuller report is also available at the cluster website (http://www.dpc.delos.info)

The Preservation Cluster has four strategic goals:

- To eliminate the duplication of effort between research activities by creating an integrating framework to co-ordinate and promote research and projects and to enable identification, collection, and sharing of knowledge and expertise
- To examine core issues that will deliver essential guidelines, methods, and tools to enable the construction of preservation functionality within digital library activities and deliverables are created.
- To establish testbeds and validation metrics. These will provide a framework for testing preservation strategies, for establishing the preservation worthiness of digital library implementations, and create greater w comparability between research and implementation activities.
- To relate the digital preservation research agenda more directly to the development of exploitable product opportunities and to develop links with the industrial sectors.

Cluster activities

The Cluster has focused its activities on five major topics.

- Establish a framework for a digital preservation testbed environment and produce metrics for testing and validating digital preservation strategies.
- Contribute to the development of digital repository frameworks and mechanisms for validating the suitability of digital repository implementations. Evaluate the current and emerging systems and storage models for digital repositories.
- Contribute to the development of file format registries and the mechanisms for their use through the definition of relationship between file format types and preservation methods and to assess the viability of producing generic metrics to measure the viability of this preservation approach.
- Define framework for documenting behaviour and functionality. Develop an overview of the attributes of functionality and behaviour that need to be represented and mechanisms for representing them.
- Develop the requirements for a preservation functionality-modeling tool and integrate that into design and development technologies.

In the near term the activities are concentrated on the following:

- integrating the preservation concepts that have been developed by the cluster with the digital library reference model that is being defined by DELOS;
- making progress on the semi-automation of the processes of ingesting material into preservation environments so as to improve construction of digital libraries;
- examining how best to deliver a substantial corpus of documents that will support measurable research in the area of automated metadata extraction;
- completing the delivery of tools to support the application of utility analysis to the selection of preservation approaches;
- examining processes of integrating the preservation tools developed by the cluster to support automatic re-appraisal of holdings.

Cluster coordinator

Dr Seamus Ross, HATII, University of Glasgow, U.K., s.ross@hatii.arts.gla.ac.uk

DELOS Digital Preservation Testbed for Testing and Evaluating Digital Preservation Solutions

Giuseppe Amato, ISTI-CNR, Giuseppe.amato@isti.cnr.it
Franca Debole, ISTI-CNR, franca.debole@isti.cnr.it
Hans Hofman, National Archives of the Netherlands, hans.hofman@nationaalarchief.nl
Max Kaiser, Austrian National Library, max.kaiser@onb.ac.at
Heike Neuroth, University of Goettingen, neuroth@mail.sub.uni-goettingen.de
Eleonora Nicchiarelli, Austrian National Library, Eleonora.nicchiarelli@onb.ac.at
Andreas Rauber, Vienna University of Technology, rauber@ifs.tuwien.ac.at
Stefan Strathmann, University of Goettingen, strathmann@mail.sub.uni-goettingen.de
Stephan Strodl, Vienna University of Technology, strodl@ifs.tuwien.ac.at

Introduction

Preservation projects have the choice between a wide array of different preservation solutions without knowing which of them best fits their requirements. The DELOS Cluster 6 on Digital Preservation developed the DELOS Digital Preservation Testbed. It provides an approach to make informed and accountable decisions on which solution to implement in order to optimally preserve digital objects for a given purpose.

Motivation

An increasing amount of our cultural and scientific heritage is being produced and maintained digitally, providing enormous benefits in terms of access, requiring less storage space and allowing for easier handling. Yet, all these enormous amounts of information are at risk of being lost due to their dependence on both current hardware and software for rendering and interaction. In order to mitigate this risk, a range of approaches for digital preservation, such as migration or emulation, to name the most prominent ones, are being investigated. For each of these approaches, a range of tools are slowly becoming available to assist in the long term preservation endeavour. Yet, each of these tools has different characteristics, performs differently, preserves different aspects of a digital object while loosing others. It is thus of eminent importance to provide assistance in the process of selecting the optimal preservation strategy for a given setting.

Testbed Framework

The DELOS Digital Preservation Testbed [1] allows institutions to evaluate preservation strategies by enforcing the precise definition of preservation requirements and supporting the documentation of the processes and experiments. It provides a means to make informed and well-documented decisions, establishing a trusted preservation process.

Figure 1 provides an overview of the workflow within the DELOS Digital Preservation Testbed. The process consists of 14 steps which are grouped into 3 stages. The first steps are the definition of the project's basic characteristics, i.e. the preservation setting. A set of representative objects is then identified, which shall be used for evaluation purposes. Next, all requirements and goals are defined, which should be fulfilled by the preservation solution. In the so-called objective tree, different goals and requirements, high-level and detailed ones, are collected and organized in a tree structure. Finally measurable units, such as Euro for costs or a subjective ranking for complexity, are assigned to the leaves of the objective tree.

In the second part of the testbed possible alternative preservation strategies are listed. Alternatives can be from all different preservation strategies, such as specific emulators, or the conversion of digital objects from one specific format to another (version of the same or a different) format, using a specific tool on a given target platform. For each alternative the resources for the evaluation experiments are selected. After that a Go/No-Go-Decision to continue, stop or redefine the process is made, based on the usefulness and cost-effectiveness of the procedure, the required resources and the expected results. If the decision is positive the required tools are set up and a set of experiments is performed. Afterwards the alternatives are evaluated with respect to the criteria defined in the first part. Here, actual measurements are collected on the objects selected for evaluation.

In the third part the evaluations per alternative are transformed to comparable numbers by using transformations tables, where the translation between the previously measured units (Euro, minutes) into a standardized scale is defined. Importance factors are assigned to explicitly describe and weight, which criteria have a major or minor impact on the final decision. Next, the comparable values are multiplied with the importance factors and finally aggregated to one final value per alternative. This single value per alternative can be used to rank the alternatives, while the results within the individual branches of the objective tree make their advantages and disadvantages in specific sub-criteria clearly visible, thus assisting in a decision on the combination of different preservation strategies.

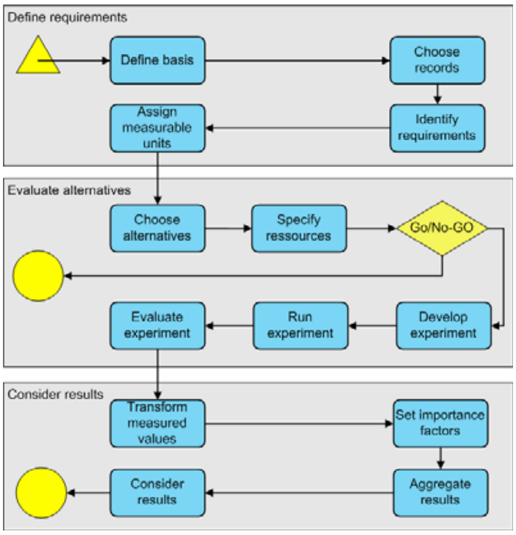


Figure 6 DP Testbed Workflow

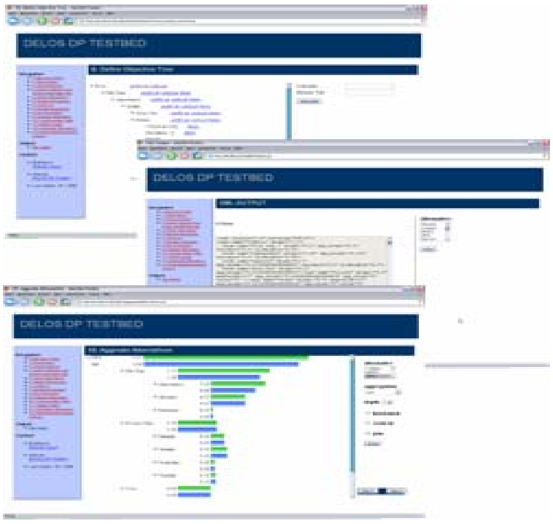


Figure 2 Screenshots of Digital Preservation software

Digital Preservation Software

The software of the DELOS Digital Preservation Testbed was developed to support the workflow and make the evaluation process easier applicable. The web based application documents each individual step in a database. Additionally the software supports the weighting of different users and an automated sensitivity analysis for the weights. The analysis determines the effect of small changes in the weight of the criteria on the final ranking of the alternatives. The DELOS Digital Preservation Testbed software implementation is available at www.ifs.tuwien.ac.at/dp. Some screenshots are shown in Figure 2.

Case Studies

Several real world case studies were performed with different partner institutions.

- Video Records of the Austrian Phonogram Archive

The Austrian Phonogram Archive is re-considering its appraisal regulations for video files, specially with respect to optimal source format standards to migrate from. A case study took place to evaluate the performance of potential migration tools and source formats.

- Document Records of the Dutch National Archive

The Dutch National Archive is responsible for storing all information, which is generated by the Dutch government, ministries and official bodies. The case study tried to define the requirements for the preservation different kinds of documents, such as video and audio document focusing particularly on the record characteristics.

- Master thesis with the Austrian National Library

A future project of the Austrian National Library is the collection of master thesis from Austrian Universities. The case study should give a starting point to identify the objectives for master thesis in context of digital preservation.

- Electronic journal with the State and University Library Göttingen

The collection of „Electronic Journal of Differential Equations" is held by the State and University Library Göttingen. The requirements and goals for the long term storage was specified for the hierarchy structured collection of annual journals.

A set of case studies are being performed to identify typical requirements and goals for prototypical institutions and different types of digital objects.

References

[1] Hofman H. and Rauber A. 2006. *Identifying, Evaluating and Selecting Preservation Methods: an Introduction to the DELOS Testbed and Utility Analysis*. Lecture given at the DELOS Summer School on Digital Preservation. San Miniato (Italy), 4-10 June 2006.

Digital Preservation Automated Ingest and Appraisal Metadata
A Progress Report on Automated Metadata Extraction

Yunhyong Kim, University of Glagow, y.kim@hatii.arts.gla.ac.uk
Seamus Ross, HATII, University of Glagow, s.ross@hatii.arts.gla.ac.uk

Progress Report

Rich descriptive, administrative, and technical metadata play a key role in the management of digital collections ([24], [14]). For example they can contribute to improving the speed and efficiency of retrieval. They play a key role in the management of digital objects within digital libraries and archives. As the DELOS/NSF ([7], [8], [9]) and PREMIS working groups ([22]) noted, when done manually, metadata are expensive to create and maintain. The manual collection of metadata can not keep pace with the number of digital objects that need to be documented. Automatic extraction of metadata would be an invaluable step in the automation of appraisal, selection, and ingest of digital material. ERPANET's Packaged Object Ingest Project ([11]) illustrated that only a limited number of automatic extraction tools for metadata are available and these are mostly geared to extracting technical metadata (e.g. DROID ([19]) and Metadata Extraction Tool ([20])). Although there are efforts to provide tools (e.g. MetadataExtractor from University of Waterloo, the Dublin Core Initiative ([10], [6]), MARS (Medical Article Record System)developed at US National Library of Medicine ([27]), Automatic Metadata Generation at the Catholic University of Leuven([1])) for extracting limited descriptive metadata (e.g. title, author and keywords) these often rely on structured documents (e.g. HTML and XML) and their precision and usefulness is constrained. For instance, and do not provide a means by which to extract sufficiently rich metadata (e.g. content summary) from either structured or unstructured digital objects. Also, we lack an automated extraction tool for high-level semantic metadata (such as content summary) appropriate for use by digital repositories; most work involving the automatic extraction of genres, subject classification and content summary lie scattered around in information extraction and language processing communities(e.g. [4], [15], [16], [23], [26], [28]).

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 last year's report we described the preliminary experimentation that we had conducted in automated metadata extraction and examined the work that had been done so far [25]. Over the past year we have made progress on refining the approach and in understanding how to constrain the search domain so that we can effectively target extraction tools. As a first step towards constructing this prototype we have focused on genre classification of documents represented in PDF ([21]). We have opted to consider sixty genres and have discussed this elsewhere [17]. This list does not represent a complete spectrum of possible genres or necessarily an optimal genre classification; it provides a baseline from which to assess what is possible. The classification is extensible. We have focused our attention on information from genres represented in PDF files. Limiting this research to one file type allowed us to bound the problem space further. We selected PDF because it is widely used, is portable, benefits from a variety of processing tools, is flexible enough to support the inclusion of different types of objects (e.g. images, links), is used to present a diversity of genre, and is a representation format in which documents are ingested in substantial quantities into digital libraries, archives, and eprint services.

Identifying the genre first limits the scope of document forms from which to extract other metadata by reducing the search space, i.e., within a single genre, metadata such as author, keywords, identification numbers or references can be expected to appear in a specific style and region. By beginning with genre classification it is possible to limit the scope of document forms from which to extract other metadata. By reducing the metadata search space metadata such as author, keywords, identification numbers or references can be predicted to appear in a specific style and region within a single genre. In fact, independent work exists on extraction of keywords, subject and summarisation within specific genre which can be combined with genre classification for metadata extraction across domains (e.g. [2], [12], [13], [26]). Resources available for extracting further metadata varies by genre; for instance, research articles unlike newspaper articles come with a list of citations closely related to the original article leading to better subject classification. Genre classification will facilitate automating the identification, selection, and acquisition of materials in keeping with local collecting policies.

In taking the genre classification forward a multi-facetted approach built on looking at documents as an object exhibiting a specific visual format (i.e. image processing), as a linear layout of strings with grammar (i.e.

language modeling), as an object with stylo-metric signatures (e.g. font style, font size, length of words), as an object with intended meaning and purpose (i.e. semantic analysis), and as an object linked to objects previously classified and other external sources (externally collected context) has been adopted.

So far experimentation has been conducted using two classifiers. First, an *Image classifier*, which depends on features extracted from the PDF document when handled as an image. It converts the first page of the PDF file to a low resolution image expressed as pixel values. This is then sectioned into ten regions for an assessment of the number of non-white pixels. Each region is rated as level 0, 1, 2, 3 with the larger number indicating a higher density of non-white space. The result is statistically modelled using the Maximum Entropy principle with MaxEnt developed by Zhang ([29]). Second we implemented a *Language model classifier*, which depends on an N-gram model on the level of words. 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 has been modelled by the BOW toolkit ([18]) using the default Naïve Bayes model without a stoplist. This has been promising [18]. As a next step forward we are going to integrate more classifiers integrating more classifiers. An *extended image classifier* could examine pages other than the just first page (as done here), or examine the image of several pages in combination: different pages may have different levels of impact on genre classification, while processing several pages in combination may provide more information.. A *language model classifier on the level of POS and phrases would use* a N-gram language model built on the Part-of-speech tags (for instance, tags denoting words as a verb, noun or preposition) of the underlying text of the document and also on partial chunks resulting from detection of phrases (e.g. noun, verb or prepositional phrases). A *stylometric classifier* taking its cue from positioning of text and image blocks, font styles, font size, length of the document, average sentence lengths and word lengths. A *semantic classifier* would combine extraction of keywords, subjective or objective noun phrases (e.g. using [23]). Finally a c*lassifier based on available external information* 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.

Making progress on the genre extractor provides the basis for constructing an efficient tool. Extension of the tool to extract author, title, date, identifier, keywords, language, summarizations, and other compositional properties can be targeted based upon genre and will, thereby, improve the precision of these other extractors. When the genre classifier is refined for PDF documents, extending it to cover other document format types (e.g. Open Office, Word, LaTeX) will be straightforward. Progress towards laying the foundation for a 'preliminary framework for designing prototype tools for assisting with preservation quality metadata extraction for ingest into digital repository' is being made by this task.

Acknowledgements

While aspects of this research are supported by *DELOS: Network of Excellence on Digital Libraries* (G038-507618) funded under the European Commission’s IST 6th Framework Programme, major funding comes from the The Digital Curation Centre (DCC). The DCC is supported by a grant from the Joint Information Systems Committee (JISC) and the e-Science Core Programme of the Engineering and Physical Sciences Research Council (EPSRC). The EPSRC supports (GR/T07374/01) the DCC’s research programme. Even with this support the scale of the research is such that we need to seek additional funding.

References

[1] Automatic Metadata Generation, http://www.cs.kuleuven.ac.be/~hmdb/amg/documentation.php

[2] Bekkerman R, McCallum A, and Huang G, 'Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora', in *CIIR Technical Report*, IR-418 (2004).

[3] Biber D, Dimensions of Register Variation: a Cross-Linguistic Comparison, Cambridge (1995).

[4]Boese E S, 'Stereotyping the web: genre classification of web documents', Master's thesis, Colorado State University (2005).

[5] Digital Curation Centre, http://www.dcc.ac.uk

[6] DC-dot, Dublin Core metadata editor, http:// www.ukoln.ac.uk/metadata/dcdot/

[7] DELOS, http://www.delos.info/

[8] NSF, http:// www.dli2.nsf.gov/intl.html

[9] 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

[10] Dublin Core Initiative, http://dublincore.org/tools/#automaticextraction

[11] ERPANET, Packaged Object Ingest Project, http://www.erpanet.org/events/2003/rome/presentations/ ross rusbridge pres.pdf

[12] Giufirida G, Shek E, and Yang J, 'Knowledgebased Metadata Extraction from PostScript File', *Proc. 5th ACM Intl. conf. Digital Libraries* (2000) 77-84.

[13] Han H, Giles L, Manavoglu E, Zha H, Zhang Z and Fox E A, 'Automatic Document Metadata Extraction using Support Vector Machines', *Proc. 3rd ACM/IEEE-CS conf. Digital Libraries* (2000) 37-48.

[14] NSF-DELOS Working Group on Digital Archiving: 'Invest to Save', Report DELOS and NSF Workgroup on Digital Preservation and Archiving (2003) http://eprints.erpanet.org/94/01/NSF_Delos_WG_Pres_final.pdf

[15] Karlgren J and Cutting D, 'Recognizing Text Genres with Simple Metric using Discriminant Analysis', *Proc. 15th conf. Comp. Ling.*, Vol 2 (1994) 1071-1075

[16] Kessler B, Nunberg G, Schuetze H, 'Automatic Detection of Text Genre', *Proc. 35th Ann. Meeting ACL* (1997) 32-38.

[17] Kim Y and Ross S, Genre Classification in Automated Ingest and Appraisal Metadata, J. Gonzalo et al. (Eds.): *ECDL 2006*, LNCS 4172, 63–74, 2006.

[18] McCallum A, Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering, http://www.cs.cmu.edu/ mccallum/ bow/ (1998)

[19] National Archives, DROID (Digital Object Identification), http://www.nationalarchives.gov.uk/

aboutapps/pronom/droid.htm

[20] National Library of New Zealand, Metadata Extraction Tool, http://www.natlib.govt.nz/en/

whatsnew/4initiatives.html#extraction

[21] Adobe Acrobat PDF specification, http://partners.adobe.com/public/developer/pdf/index_reference.html

[22] PREMIS Working Group, http://www.oclc.org/research/projects/pmwg/

[23] Riloff E, Wiebe J, and Wilson T, `Learning Subjective Nouns using Extraction Pattern Bootstrapping', *Proc. 7th CoNLL*, (2003) 25-32

[24] Ross S and Hedstrom M, 'Preservation Research and Sustainable Digital Libraries', *Int Journal of Digital Libraries* (Springer) (2005) DOI: 10.1007/s00799-004-0099-3.

[25] Ross S and Kim Y, 2005, 'Digital Preservation Automated Ingest and Appraisal Metadata', in C Thanos (ed), *DELOS Research Activities 2005*, Pisa: ISTI-CNR, 60-61, ISBN 2-912335-14-0

[26] Sebastiani F, 'Machine Learning in Automated Text Categorization', *ACM Computing Surveys*, Vol. 34 (2002) 1-47.

[27] 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.

[28] Witte R, Krestel R, and Bergler S, 'ERSS 2005:Coreference-based Summarization Reloaded', *DUC 2005 Document Understanding Workshop*.

[29] Zhang L, Maximum Entropy Toolkit for Python and C++, LGPL license, http:// homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html

Investigation of Automation of Re-Appraisal and an Open Testbed Document Corpus

Seamus Ross, HATII, University of Glagow, s.ross@hatii.arts.gla.ac.uk

As Tasks 6.8 and 6.9 have only just begun under JPA3 this report is very preliminary, but we felt that it was essential that we should share information about our objectives with a broader community to promote greater engagement in our work.

The Cluster as a whole sees itself as integrating the work conducted under WP6-JPA1 and WP6-JPA2 into a single stream which will allow us to take the results obtain so far and extend them to address a key preservation problem: appraisal and selection of content for selection and ingest into digital libraries. In an effort to approach this problem the Cluster has begun two new tasks:

- Task 6.8: Investigation of the Automation of Appraisal.
- Task 6.9: Development of a Open Testbed Document Corpus

Task 6.8: Investigation of the Automation of Appraisal

The usefulness of digital libraries to the archival community will be influenced by our ability to automate the process of appraisal and re-appraisal of holdings within a digital library at different points during the lifecycle of a digital object. Building on work completed in JPA1 Task 6.4, JPA2 Tasks 6.5, 6.6, and 6.7 this new task is investigating how the process of automatic re-appraisal (resulting in either the disposal or retention) of digital holdings might be effectively handled in the context of digital libraries. Organisations repeatedly take decisions on what information (or digital objects) should be preserved and for how long; they retake these decisions repeatable about the same materials as they come up for re-appraisal. In the case of records management and archiving criteria governing what to keep and what to discard are usually based upon such factors as organisational needs/objectives, juridical requirements, and information value that are relevant to the business context of the organization (whether a library, a public sector institution or a commercial company). This is happening in government organizations, business companies and memory organisations including digital libraries. The main reason for this is that preserving too much digital material makes it difficult to manage it cost effectively; so less may prove to be more. In the paper world this appraisal and selection process is conducted manually and requires an enormous amount of effort. In a digital environment new approaches are possible, one of them could be automating parts of the actual appraisal and selection process. Appraisal points arise at the time of each preservation action, such as ingest, migration, and even access. Based on these stages it will be possible to identify where automation may best support the process.

Crucial in the processes of appraisal are an understanding of the material that is being appraised, adequate metadata to support the appraisal process, and tools to test preservation scenarios. Our modelling work builds on the efforts to extract information about the digital objects (either about content or context or both as addressed in Tasks 6.6 and 6.7). The new core work involves analysing and modelling appraisal criteria, rules and processes, and experimenting with their automation. Crucial in this process will be a better understanding of the role of preservation metadata in the appraisal, selection and disposal of digital objects. As has been noted by the PREMIS working group (2004) our understanding of the useful metadata for digital preservation is limited and no studies have effectively demonstrated the value metadata to preservation processes.

As a result of earlier research within WP6 we have some building blocks to allow us to experiment with such automated activities. The extraction tool(s) provide excellent mechanism/method for extracting information about the content, the context of origin as well as the technical nature of a digital object or any aggregation of objects. Using knowledge representation and processing techniques we intend to automate the appraisal decision making processes, doing this depends on appraisal modeling and the integrating the classes of information created through the work of Task6.6 and Task 6.7.

Task 6.9 – Development of a Open Testbed Document Corpus.

The quantities of digital materials to be ingested into digital libraries make it impossible for each of them to be described manually. While for many user communities commonly available search techniques are sufficient to mine and identify documents, for some knowledge about such characteristics of a document as context, subject, technical characteristics, and genre are critical to ensuring that we can compare different metadata extraction approaches. In work on Task 6.7 we recognised that to do any meaningful work we needed a substantial, consistently designed and documented corpus. The process of working on appraisal as described in Task 6.8 also requires a corpus with which we can experiment to test the implications of automated appraisal decisions. We looked at how a document corpus might be generated and concluded that the design and implementation of such a corpus needed serious attention. Such factors as scale, detail of labelling, distribution of file types and technical characteristics needed to be analysed in more detail to ensure that the corpus has maximum advantage to the preservation research community for tool testing, tool training, and comparative evaluation and benchmarking of methods and tools.

The DELOS test bed (designed under JPA1) will provide us with mechanism to experiment with any outputs from Task 6.8 and the document corpus from Task 6.9 will provide a core tool in this process. We envisaged for instance that experiments might enable us to test different approaches to appraisal such as identifying the selection (or retention) rules and subsequently the evaluation of the value of information or digital objects prioritised by them. Evaluation can be based on different methods, such as analyzing the context in which the digital objects were created (e.g. the business activities) or analyzing their content. Selection which includes the attribution of relative or absolute values to the digital objects could be pursued, and the effect of disposal, which is the actual application of the appraisal and selection decisions, on the value of the corpus could be tested. These experiments would be in addition to comparative validation of automated metadata extraction methods.

The key cluster participants, Michael Day (UKOLN, University of Bath, UK), Mariella Guercio (University of Urbino, Italy), Hans Hofman (Nationaal Archief, Netherlands), Yunghyong Kim (University of Glasgow, UK), Andreas Rauber (Vienna University of Technology, Austria), Seamus Ross (HATII University of Glasgow), Stefan Strathmann (Goettingen State and University Library, Germany), Manfred Thaller (University at Cologne, Germany), and colleagues at their institutions are working together to pull these strands together in JPA3 and more tightly in JPA4. Automation of the process of selection and appraisal is core to ensuring incorporation of preservation methodologies within digital libraries.

Evaluation

Cluster objectives

Digital libraries need to be evaluated as systems and as services to determine how useful, usable, and economical they are and whether they achieve reasonable cost-benefit ratios. Results of evaluation studies can provide strategic guidance for the design and deployment of future systems, can assist in determining whether digital libraries address the appropriate social, cultural, and economic problems, and whether they are as maintainable as possible. Consistent evaluation methods also will enable comparison between systems and services.

The evaluation cluster is working both on evaluation methodologies in general as well as on providing the infrastructure for specific evaluations. Thus, the following objectives are addressed:

- *Development of a comprehensive theoretical framework for DL evaluation,* which can serve as reference point for evaluation studies in the DL area.
- *Research on new methodologies,* in order to overcome the lack of appropriate evaluation approaches and methods.
- *Development of toolkits and test-beds* in order to enable new evaluations and to ease the application of standard evaluation methods.

Cluster activities

In order to reach these goals, the following activities are being carried out:

- *Workshops on DL evaluation,* for collecting existing evaluation approaches and methods.
- *Evaluation support to the DL community,* by creating an evaluation forum for enabling communication between evaluation specialists and DL developers.
- Development of new approaches and methods, in order to overcome the weaknesses of current approaches and the lack of methods for new types of applications.
- Development of evaluation toolkits, e.g. for collecting and analyzing experimental data.
- Creation of test-beds for new content and usage types in DLs, by starting from the existing test-beds for XML and cross-lingual retrieval and extending these towards new media, applications and usage types.
- Creation of test-beds for usage-oriented evaluation, by extending existing test-beds or by creation of test-beds of user interactions

Cluster coordinator

Norbert Fuhr, Universität Duisburg-Essen, Germany, fuhr@uni-duisburg.de

The INEX Initiative for the Evaluation of XML Document Access and Retrieval

Mounia Lalmas, Queen Mary University of London, mounia@dcs.qmul.ac.uk
Anastasios Tombros, Queen Mary University of London, tassos@dcs.qmul.ac.uk

Keywords: H.3.7 [Information Storage & Retrieval]:Digital Libraries; H.3.3 [Information Storage & Retrieval]:Information Search & Retrieval-Search Process; H.3.4 [Information Storage & Retrieval]:Systems and Software-Performance evaluation.

Initiative for the Evaluation of XML Retrieval

The amount of information accessible has transformed the Web into a universal public information repository. A major outcome of this transformation has been, and still remains, the promotion of knowledge sharing. This has forced traditional information providers, like libraries, to also publish their information on the Web. However the fact that the Web is growing at a phenomenal rate makes it difficult to effectively access all the published information. One reason is that this information is mostly published using HTML, a markup language that cannot accurately describe a page's content and structure. Therefore, modern Web applications, like digital libraries, have been increasingly publishing their information using the eXtensible Markup Language (XML) in order to bring some order to the Web.

The continuous growth in XML information repositories has been matched by increasing efforts in the development of XML retrieval systems, in large part aiming at supporting content-oriented XML retrieval. These systems exploit the available structural information, as marked up in XML, in documents, in order to implement a more focussed retrieval strategy and return document components - the so-called XML elements - instead of complete documents in response to a user query. This focussed retrieval approach is of particular benefit for repositories containing long documents or documents covering a wide variety of topics (e.g. books, user manuals, legal documents), where users' effort to locate relevant content can be reduced by directing them to the most relevant parts of these documents.

The DELOS task 7.3 is concerned with the evaluation of content-oriented access to XML documents. The provision of effective access to XML-based content has become a key research issue, and is the focal point of XML retrieval research. Evaluating how good these systems are, hence, requires test-beds where the evaluation paradigms are provided according to criteria that take into account the imposed structural aspects.

In 2002, the Initiative for the Evaluation of XML Retrieval (INEX)[6] started to address these issues. INEX has a strong international character; participants from over 70 organisations, distributed across Europe, North America, Australia, New Zealand and Asia, have participated in this year's fifth INEX run. The aim of the INEX initiative is to establish an infrastructure and to provide means, in the form of a large XML test collection and appropriate scoring methods, for the evaluation of content-oriented XML retrieval systems. As for each round of INEX, INEX 2006 started in late March and will end with a workshop in December in Dagsthul, Germany.

Main INEX Activities

Ad-hoc retrieval task. The main retrieval task in INEX is the ad-hoc retrieval of XML documents. In IR literature, ad-hoc retrieval is described as a simulation of how a library might be used, and it involves the searching of a static set of documents using a new set of topics (queries). While the principle is the same, the difference for INEX is that the library consists of XML documents, the queries may contain both content and structural conditions and, in response to a query, arbitrary XML elements may be retrieved from the library.

Within the main ad-hoc retrieval task, two sub-tasks have been identified depending on how structural constraints are expressed in queries: a) in the *Content-Only (**CO**)* sub-task, queries ignore the document structure and contain only content-related conditions, b) in the ***+S** sub-task* (an extension of the CO task), a user may decide to add structural hints to his query to narrow down the number of returned documents resulting from a CO query. For each of the above two sub-tasks, four specific retrieval strategies are investigated.

[6] http://inex.is.informatik.uni-duisburg.de/

In a *focussed* strategy, it is assumed that a user prefers a single element that most exhaustively discusses the topic of the query (most exhaustive element), while at the same time it is most specific only to that topic (most specific element). In a *thorough* strategy, a user is assumed to prefer all highly exhaustive and specific elements. In a *relevant in context* strategy, systems return (all) relevant elements grouped by article. For each article, an unranked set of elements is returned, covering the relevant material in the article. Overlap is not permitted in this strategy (i.e. a section and one of its paragraphs can not be returned as separate elements). In the *best entry point in context* strategy, the aim is to find the best-entry-point for starting to read articles with relevance. As a result, even an article completely devoted to the topic of request will only have one best starting point to read.

Additional tasks. In addition to the evaluation of retrieval effectiveness for the ad-hoc task, further research issues, with direct applications to digital libraries, are being explored in INEX in the form of research tracks. The current run of INEX, which started in April 2006, includes a number of additional research tracks. Six of the tracks (relevance feedback, natural language processing, heterogeneous collection, interactive, multimedia, XML document mining) were also included in INEX 2005 [2]. Two additional tasks are part of INEX 2006.

User-case studies track. This track is an experiment into how an XML search engine will be used. Participants are asked to hypothesise how XML retrieval might be used (e.g. why might users prefer this technology over other technologies, what is the work-task the user has, how do users interact with elements, etc.). The main aim of this track is to provide a better understanding of who the users are, and how they might use XML retrieval.

XML Entity Ranking track. The Expert Search task in the 2005 TREC Enterprise Track has evaluated systems that return a list of entities (people's names) who are knowledgeable about a certain topic (e.g., "information retrieval"). The idea of the entity ranking track is to generalise this setting to arbitrary entity types. Consider for example a Famous Actor task. Given a topic "1930s", Astaire, Chaplin, Gable and Garbo should be returned, whereas a topic "action" should result in Schwarzenegger, Stallone and Van Damme. A setting with semi-structured data seems particularly suited as a basis for such a system, which could use the text elements, but also structural and linking information. The primary interest is not to address the entity extraction part of the problem, but really how to associate entities to a topic text.

Test Collection Building

Document collections. Evaluating the effectiveness of XML retrieval systems requires a collection of documents marked-up in XML. INEX created an XML test collection consisting of publications donated by the IEEE Computer Society, referred to as the INEX IEEE collection. The collection consisted of the full-text of 16,819 articles, marked-up in XML, from 12 magazines and 6 transactions of the IEEE Computer Society's publications, covering the period of 1995-2004, and totalling 754 MB in size, and 11 millions in number of elements. On average, an article contained 1,532 XML nodes, where the average depth of the node is 6.9.

INEX 2006 uses a new document collection, which is made from English documents from Wikipedia[7][1]. The collection consists of the full-texts, marked-up in XML, of 659,388 articles of the Wikipedia project, covering a hierarchy of 113,483 categories (articles in Wikipedia are organised into categories), and totalling more than 60 GB (4.6 GB without images). The collection has a structure similar to the INEX IEEE collection; on average, an article contains 161.35 XML nodes, where the average depth of an element is 6.72.

Topics. In previous years, different topic types have been used for the ad-hoc retrieval task at INEX [2] (i.e., a distinction was made between Content Only (*CO*) and Content and Structure (*CAS*) topics). Following initial trials at INEX 2005 and recommendations from the 2005 Dagstuhl workshop, these topic types are all merged into one type for INEX 2006: Content Only + Structure (CO+S). In the 2006 topics all the information needed by the different ad-hoc tasks and other tracks are expressed in the individual topic parts, and only one topic type is therefore needed. The 2006 CO+S topics consist of a number of parts (e.g. <title>, <description>, <narrative>, <ontopic-keywords>, etc.) that contain information at different granularity levels about the topic (e.g. a title contains a short description of the information need, where the narrative contains a detailed explanation of the information need and the description of what makes an element relevant or not). A total of 201 topics were submitted by the participating organisations, of which 125 have been selected as the final topics.

Conclusions

With the inclusion of the various research tracks, INEX is expanding in scope and in the number of participating organisations. INEX has also acquired new collections of XML documents in an effort to enhance the evaluation

[7] http://en.wikipedia.org

environment. INEX has shown that XML retrieval is a challenging field within IR and DL research. In addition to learning more about XML retrieval approaches, INEX is making steps in the evaluation methodology for accessing and retrieving XML documents.

References

[1] Denoyer L. and Gallinari P. 2006. The Wikipedia XML corpus. In *SIGIR Forum* 40(1): 64-69.

[2] Fuhr N., Lalmas M., Malik S. and Szlavik Z. (eds.) 2005. Advances in XML Information Retrieval. Revised Selected Papers from the 3rd International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2004). Schloss Dagstuhl (Germany), 6-8 December 2004. Lecture Notes in Computer Science 3493. Berlin-Heidelberg: Springer.

Multilingual Information Access for Digital Libraries: The Impact of Evaluation on System Development

Carol Peters, ISTI-CNR, carol.peters@isti.cnr.it

Categories, Subject descriptors and Keywords

H.3.3 [Information Storage and Retrieval]: Search process; H.3.4 [Systems and Software]: performance evaluation, Multingual information access, System evaluation, Test collections.

Introduction

As has been made very evident in the recent press releases from the European Commission on the i2010 Digital Libraries Initiative, Multilingual Information Access (MLIA) is a key issue for European digital libraries. For this reason, the DELOS Network of Excellence, under both FP5 and FP6, has supported the Cross Language Evaluation Forum (CLEF) since it began activities in 2000. Here below we provide a very brief summary of the CLEF activities and the impact that these have had on R&D in the MLIA domain.. For further information and access to a wide literature on the topic, see the CLEF website: http://www.clef-campaign.org.

CLEF Activities

The objective of CLEF is to promote research in the field of multilingual system development. This is done through the organisation of annual evaluation campaigns in which a series of tracks designed to test different aspects of mono- and cross-language information retrieval (IR) are offered. The intention is to encourage experimentation with all kinds of multilingual information access – from the development of systems for monolingual retrieval operating on many languages to the implementation of complete multilingual multimedia search services. This has been achieved by offering an increasingly complex and varied set of evaluation tasks over the years. The aim is not only to meet but also to anticipate the emerging needs of the R&D community and to encourage the development of next generation multilingual IR systems.

Figures 1 and 2 show the increase in tracks at CLEF over the years and the number of participating groups per track (groups can participate in more than one track). Most tracks include a number of different tasks which vary campaign by campaign both in scope and languages involved. Thus, although CLEF 2006 has the same eight tracks as CLEF 2005, many of the tasks offered are new. In particular, this year, the Ad Hoc task has focussed on testing of cross-language systems for lesser studied languages (including Amharic, Hindi, Oromo and Telugu) and on failure analysis with a “robust task” which uses test collections from previous years. The interactive track is using data from the Flikr online photo archive to capture the interplay between image search and the multilingual reality of the internet [1]. The QA track has introduced a real-time task to measure not just system performance but also response times, and a task based on data from Wikipedia. ImageCLEF is using a new collection of travel photos as well as extended versions of the medical image databases used in 2005, and the cross-language speech track includes a collection of oral interviews in Czech this year. 110 groups have registered to participate in CLEF 2006 and the results will be presented at the annual workshop to be held in Alicante, Spain, in conjunction with ECDL 2006.

CLEF 2000

- mono-, bi- and multilingual textual document retrieval (Ad Hoc)

- mono- and cross-language information on structured scientific data (Domain-Specific)

CLEF 2001

- interactive cross-language retrieval (iCLEF)

CLEF 2002

- cross-language speech retrieval (CL-SR)

CLEF 2003

- multiple language question answering (QA@CLEF)

- cross-language retrieval in image collections (ImageCLEF)

CLEF 2005

- multilingual retrieval of Web documents (WebCLEF)

- cross-language geographical retrieval (GeoCLEF)

Table 1: CLEF 2000 – 2005: increase in tracks

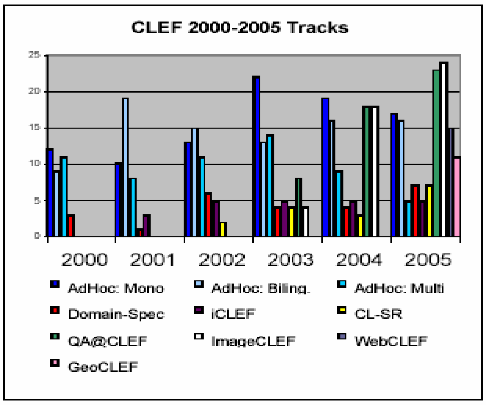


Figure 1: CLEF 2000-2005 Shift in focus

CLEF Test Collections

CLEF campaigns adopt a comparative evaluation approach in which system performance is measured using appropriate test suites [2]. These consist of sets of sample query statements often called "topics", document collections, and relevance judgments determining the set of relevant documents in a collection for a given query statement. The document collections were used to build the test sets in CLEF 2006 include:

- a multilingual comparable corpus of more than 2 million news docs in twelve European languages
- social science databases in English, German and Russian
- an archive of photos related to tourism with captions in English and German
- the LTU Tech Object Dataset, a fully categorised collection for automatic annotation task
- radiological medical databases with case notes in French, English and German
- a collection of spontaneous conversational speech derived from the Shoah archives in English and Czech
- a multilingual collection of about 2M web pages crawled from European governmental sites.

For each collection, appropriate sets of search requests and associated relevance assessments have been built. These test suites form extremely valuable and reusable resources.

Impact of CLEF

The results achieved by CLEF in these seven years of activity can be summarised in the following main points:

- Encouragement of system testing for languages other than English: CLEF target collections have been built in more than 20 different European languages.
- Development of and evaluation infrastructure (the DIRECT system) that supports the production, maintenance, enrichment and interpretation of system data for subsequent in-depth evaluation studies [3]
- Promotion of fully multilingual retrieval systems: CLEF 2003 included a task which entailed searching a collection in 8 languages, selected to cover a range of language typologies and linguistic features.
- Documented improvement in system performance for cross-language text retrieval systems: cross-language results are now regularly at 85-90% of monolingual retrieval (around 50% in 2000) [4].

- Quantitative and qualitative evidence with respect to best practice in cross-language system development and qualitative and quantitative evidence as to which methods give the best results in certain key areas, such as multilingual indexing, query translation, resolution of translation ambiguity, results merging [5].
- R&D activity in new areas such as cross-language question answering, multilingual retrieval for mixed media, and cross-language geographic information retrieval [6].
- The creation of a large set of empirical data about multilingual information access from the user perspective;
- Creation of important, reusable test collections for system benchmarking: ELDA (www.elda.org) is about to release a first CLEF test suite on its 2006 catalog.
- Building of a strong, multidisciplinary research community, see for example [6].

However, although CLEF has done much to promote the development of multilingual IR systems, so far the focus has been on building and testing research prototypes rather than developing fully operational systems. There is still a considerable gap between the research and the application communities and, despite the strong demand for and interest in multilingual IR functionality, there are still very few commercially viable systems on offer. The challenge that CLEF must face in the near future is how to best transfer the research results to the market place. Two members of the CLEF coordinating group (ISTI-CNR and UniPD) are currently collaborating in a feasibility study aimed at identifying the main issues involved in implementing full multilingual information access functionality in TEL - The European Library, within the context of the DELOS-TEL collaboration.

Project Participants

University of Padua: Maristella Agosti, Giorgio Di Nunzio, Nicola Ferro

Swedish Institute of Computer Science – SICS: Preben Hansen, Jussi Karlgren

References

[1] Karlgren J., Clough P. and Gonzalo J. 2006. Multilingual interactive experiments with Flickr. In *ERCIM News* 66: 36-37.

[2] Cleverdon C. 1977. The Cranfield tests on index language devices. In K. Sparck-Jones and P. Willett (eds.) Readings in Information Retrieval. San Francisco: Morgan Kaufmann. 47-59.

[3] Agosti M., Di Nunzio G.M. and Ferro N. 2006. A data curation approach to support in-depth evaluation studies. In F. Gey, N. Kando and C. Peters (eds), *Proceedings of the SIGIR 2006 Workshop on New Directions in Multilingual Information Access. Seattle (USA), 10 August 2006*. New York: ACM Press.

[4] Gonzalo J. and Peters C. 2005. The impact of evaluation on multilingual text retrieval. In R.A. Baeza-Yates, N. Ziviani, Gary Marchionini, Alistair Moffat, John Tait (eds.), *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005). Salvador (Brazil), 15-19 August 2005*. New York: ACM Press. 603-604.

[5] Braschler M. and Peters C. 2004. Cross-language evaluation forum: objectives, results, achievements. In *Information Retrieval* 7(1-2): 7-31.

[6] Peters C., Clough P.D., Gonzalo J., Jones G., Kluck M. and Magnini B. (eds.) 2004. Multilingual Information Access for Text, Speech and Images. Revised Selected Papers from the 5th Workshop of the Cross-Language Evaluation Forum (CLEF 2004). Bath (UK), 15-17 September 2004. Lecture Notes in Computer Science 3491. Berlin-Heidelberg: Springer.

[7] Braschler M., Ferro N. and Verleyen J. 2006. Implementing MLIA in an existing DL system. In F. Gey, N. Kando and C. Peters (eds), *Proceedings of the SIGIR 2006 Workshop on New Directions in Multilingual Information Access. Seattle (USA), 10 August 2006*.

A Digital Library Testbed Framework for the Evaluation of Architectures, Services and Execution Dynamics

Norbert Fuhr, University of Duisburg, fuhr@uni-duisburg.de
Hanne Albrechtsen, Institute of Knowledge Sharing, hanne.albrechtsen@knowshare.dk
Claus-Peter Klas, University of Duisburg-Essen, klas@uni-duisburg.de
Sarantos Kapidakis, Ionian University, sarantos@ionio.gr
Andras Micsik, Hungarian Academy of Sciences, micsik@sztaki.hu

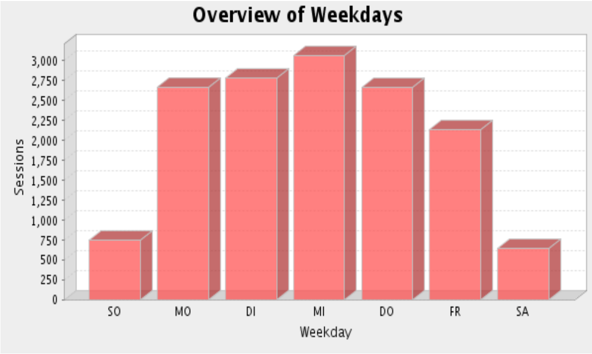
Keywords: H.3.7 [Digital Libraries]: Dissemination – System Issues – User Issues - Evaluation.

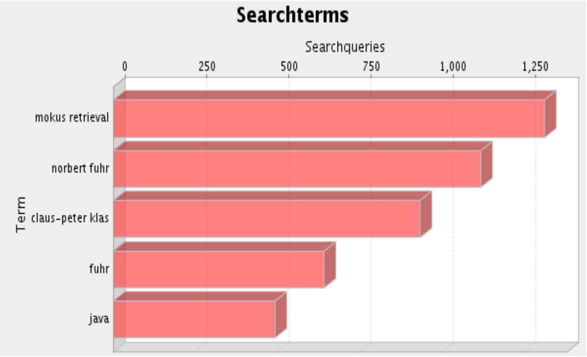
Objectives

Today most digital library (DL) evaluations use specific systems, which are difficult to compare. The aim of this effort is to provide a standard testbed framework for comparative evaluation of DL systems. Based on a theoretical framework for DL evaluation, we will develop a framework system to *guide* the stakeholders of a DL evaluation and to *provide* a communication and sharing platform. Thus, stakeholders can work more efficiently by using the DELOS evaluation framework for their scientific research. A second aim of this task is to provide actual evaluations of any aspect in the field of DLs, in order to enhance tasks and services.

Standard Logging Framework

Within the standard evaluation framework we developed the standard logging framework which is presented in full length in Klas et al, 2006. Along with the logged events we also implemented graphical tools to generate statistical information based on the logged data. Typical aggregated information can now be extracted like usage times, most common objects, etc. The example aggregations shown in Figures below were produced with the help of these tools. The statistics can be filtered by different attributes like time, (anonymous) users or user groups.





Activities

Evaluation of „The European Library“

The task 7.5 is currently in preparation of a comparative analytical and empirical evaluation of “The European Library”. The goal of the analytical evaluation is to assess the functional similarities and differences between the existing TEL search tool and the Daffodil-based search interface. The goal of the empirical evaluation is to evaluate how well each tool supports the users’ needs.

The analytical evaluation considers the following use-oriented characteristics of the two search tools:

- Usability
- Search and browse functions
- Display functions
- Feedback/help functions

The analytical evaluation will be conducted through expert walkthrough of systems/prototypes as well as a desk study of the corresponding design documentation. The expert walkthrough of the systems/prototypes will access similar information sources.

The empirical evaluation will be based on a user-centered and qualitative approach. Its focus is on the users experience with the tools. It considers the following perspectives:

- User characteristics, preferences and strategies (i.e. their experience),
- Types of activities that users carry out/tasks they perform,
- The environment in which the search tool is used, either in the natural work setting or in a controlled laboratory setting.

The empirical evaluation will be conducted in natural settings as well as in laboratory settings, and enroll representatives of two major categories of users, the stakeholders of TEL (e.g., librarians) and end-users of TEL.

On the conceptual side, group members setup an evaluation plan for a comparative evaluation. The schedule consists of several steps:

1. Pre-Evaluation of the current TEL interface according to proposed evaluation model. The goal of this pre-evaluation is to set the baseline, to capture the state of the interface.
2. Pre-Evaluation of the current Daffodil interface with respect to the TEL interface.
3. Write a proposal for the actual qualitative evaluation in contact with TEL as discussion base.
4. Prepare and execute the actual evaluation.

Evaluation of the Daffodil/DiLAS/MadCow prototype on Annotation for Collaboration.

In cooperation with group members of task 4.10 and task 4.1, group members of task 7.5 conducted an expert evaluation of the Daffodil/DiLAS/MadCow prototype, in preparation for empirical evaluation of the prototype by end-users. The expert evaluation was carried out at a meeting hosted by the Institute of Knowledge Sharing in Copenhagen, Denmark, from 22-23 May 2006. The group has produced a working paper on short term requirements for modification of the current version of Daffodil/DiLAS/MadCow (Albrechtsen et al, 2006). The focus of the requirements list in the working paper is on user experience with the Daffodil/DiLAS/MadCow annotation tool. In addition the requirements list includes some observations that may inspire a technical test of the annotation tool. Evaluation methods applied were cognitive walkthrough (general usability of the tool), participatory group evaluation (the developer and the evaluators in dialog), and collaborative work task evaluation (based on common task for the expert evaluators). The full list of requirements will be described in the forthcoming document "Evaluation of Multimedia Annotations for User Collaboration", which will be distributed to the participants of task 4.1, task 4.10 and task 7.5.

INEX: Evaluation collaboration with task 7.3

Within the INEX initiative, two tracks are in planning and under implementation based on the Daffodil framework. Within the interactive track, a snapshot of the Wikipedia is currently integrated as data content. Along with the content, the search interface has to be adopted to present the new data. Also other tools like the personal library will be adopted and used in the upcoming track.

The second track will be the interface track. Here it is planned to propose the Daffodil graphical framework as basis for evaluation of search interfaces, along with other interfaces. Each participant implements or reuses an existing search interface, either based on own frameworks (like e. g. web interfaces), or using the Daffodil system.

Task Members

University of Duisburg-Essen : Norbert Fuhr fuhr@uni-duisburg.de, Claus-Peter Klas klas@uni-duisburg.de

Institute of Knowledge Sharing: Hanne Albrechtsen hanne.albrechtsen@knowshare.dk

MTA SZTAKI DSD: Laszlo Kovacs laszlo.kovacs@sztaki.hu, Andras Micsik micsik@dsd.sztaki.hu

Ionian University: Sarantos Kapidakis sarantos@ionio.gr

References

[1] Klas C.-P., Albrechtsen H., Fuhr N., P. Hansen, E. Jacob, Kapidakis S., Kovacs L., Kriewel S., Micsik A., Papatheodorou Ch. and G. Tsakonas forthcoming. A Logging scheme for comparative digital library evaluation. In *Proceedings of the 10th European Conference on Digital Libraries (ECDL 2006). Alicante (Spain), 17-22 September 2006*.

[2] Albrechtsen H., Hansen P. and Pejtersen A.M. 2006. Evaluation of multimedia annotations for user collaboration: short term requirements for the DilAS/Daffodil/MadCow prototype 1, based on expert evaluation. Working paper (12 June 2006).

[3] Malik S., C.-P. Klas, Fuhr N., Larsen B. and Tombros A. 2006. Designing a user interface for interactive retrieval of structured documents - Lessons learned from the INEX interactive track. In *Proceedings of the 10th European Conference on Digital Libraries (ECDL 2006). Alicante (Spain), 17-22 September 2006*.

Technology Transfer

Objectives and activities

Following the terminology of the EU, all the activities related to visibility, dissemination, education and technology transfer are gathered under the umbrella “Spreading of Excellence”. The Spreading of Excellence activities fall therefore into different categories, depending on the main objective of a specific activity, its contents and the expected recipients.

For **Scientific Dissemination**, DELOS organizes both *Thematic Workshops,* intended to provide the opportunity to European researchers to present results of on-going research activities and to exchange opinions and experiences in an informal and friendly environment, and *Brainstorming Workshops,* intended to bring together top researchers from all parts of the world, focusing on specific technology issues (e.g. search engines) to produce reports on "future research direction" in the field of Digital Libraries, which can provide input for the definition of future research programmes both to the EC and to national research funding agencies.

For **Education and Training**, DELOS organizes a series of *Summer Schools*, intended to provide high-level courses on the domain of Digital Libraries and its underlying technologies. The schools are directed to members of the research community (in the wide sense): primarily graduate students, but also young researchers and professionals involved in R&D in Digital Library related areas.
DELOS organizes also a number of *Awareness Events* (at the National or Regional level). These events (workshops, courses, tutorials, demonstrations, etc) are usually organized in collaboration with application and industrial communities. Finally, DELOS is continuing the *Research Exchange Program*, which supports the visit of a young researchers from a research organization (not necessarily a DELOS member) to a DELOS member, for a period ranging from a few weeks to a few months The exchange of researchers between organizations working on joint projects or closely related topics has proven to be one of the most effective ways to achieve integration of teams, exchange of skills and results, and training to young researchers.

For **Visibility**, the main vehicle is obviously the *Web site* (www.delos.info), which provides information about the activities and events of DELOS, and also news and links about activities related to Digital Libraries going on worldwide: The main feature of the site is the access to the Delos Digital Library, which contains publications produced by DELOS partners, and provides some initial advanced functionality (such as the definition of virtual collections, selective access control, sophisticated search function). DELOS is publishing also an *Electronic Newsletters*, providing information and reports on DELOS events and on related activities of interest to the digital library community. DELOS is also supporting the *DELOS Award*, named “DELOS Research Exchange Award”, to be given each year to a young researcher who has authored a paper presented at the ECDL Conference. The award will recognize the achievement of the young researcher by offering the author a chance to spend a period of time (at least one month) at a DELOS European partner research organisation.

Finally, in the area of **Technology Transfer**, in addition to using the national or regional Awareness Events, in the last year DELOS has concentrated its outreach efforts in the library and cultural sectors, establishing good links with TEL (The European Library) and with ELAG (European Library Automation Group), and strengthening its links with MINERVA and MICHAEL. In particular, the cooperation with TEL has been structured into four “technology transfer” projects, related to functionality areas of primary relevance for TEL, namely architecture, multilinguality, personalization, user interface and visualization. The main effort will be primarily focused on the integration of DELOS-provided software and prototypes into the existing European Library system. In the following pages there is a brief description of the projects related respectively to multilinguality, personalization and user interface and visualization. The project on architecture, which includes the cooperation with TEL, has been described in the Architecture cluster (A Reference Model for Digital Libraries).

Cluster coordinator

Vittore Casarosa, ISTI-CNR, Italy, vittore.casarosa@isti.cnr.it

A Study on how to Enhance TEL with Multilingual Information Access

Maristella Agosti, University of Padua, agosti@dei.unipd.it
Martin Braschler, Zurich University of Applied Sciences Winterthur, martin.braschler@zhwin.ch
Nicola Ferro, University of Padua, ferro@dei.unipd.it

This work reports on the study conducted in collaboration between DELOS and *The European Library* (TEL)[8], a service fully funded by the participant national libraries members of the *Conference of European National Librarians* (CENL)[9], which aims at providing a co-operative framework for integrated access to the major collections of the European national libraries.

The aim of the collaboration, started in April 2006, is to conduct a feasibility study which identifies the main issues involved in implementing full *MultiLingual Information Access* (MLIA) functionalities in TEL. By full MLIA we mean the possibility for users of TEL to access and search the federated libraries in their own (or preferred) language, retrieve documents in other languages, have the results presented in an interpretable fashion (e.g. possibly with a summary of the contents in their chosen language).

The TEL project aims at providing a "low barrier of entry" in the TEL system to the national libraries which want to join it. This easiness of integration is achieved by extensively using the *Search/Retrieve via URL* (SRU)[10] protocol in order to search and retrieve documents from national libraries. In this way, the user client can be a simple browser, which exploits SRU as a means for uniformly accessing national libraries. With this objective in mind, TEL is constituted by three components:

- a Web server, which provides users with the TEL portal;
- a central index, which harvests catalogue records from national libraries which support *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH) and provides integrated access to them via SRU;
- a gateway between SRU and Z39.50: it allows national libraries which support only Z39.50[11] to be accessible via SRU.

The architecture and functioning of the TEL system pose some problems when planning to introduce MLIA. The TEL system has no control on queries sent to the national libraries, since the client browser directly manages the interaction with national library systems via SRU. As a consequence, introducing MLIA functionalities into the TEL system would have no effect on the national library systems. Thus, in order to achieve full MLIA functionalities, not only the TEL system but also all the national library systems should be modified and this is an unviable option that would require a very big effort and disregards the "low barrier of entry" guideline adopted in designing the TEL system.

In order to avoid the problem discussed above, still offering some MLIA functionalities, we are considering to introduce an *isolated query translation* step in the query processing. In addition, the TEL central index harvests catalogue records from national libraries, which beside catalogue metadata may contain other information useful for applying MLIA techniques, such as an abstract. Since the central index is completely under the control of the TEL system, we are considering to extend its functionalities by adding a component able to translate the catalogue records in order to perform MLIA on them. We call this approach *pseudo-translation*.

[8] http://www.theeuropeanlibrary.org/

[9] http://www.cenl.org/

[10] http://www.loc.gov/standards/sru/

[11] http://www.loc.gov/z3950/agency/

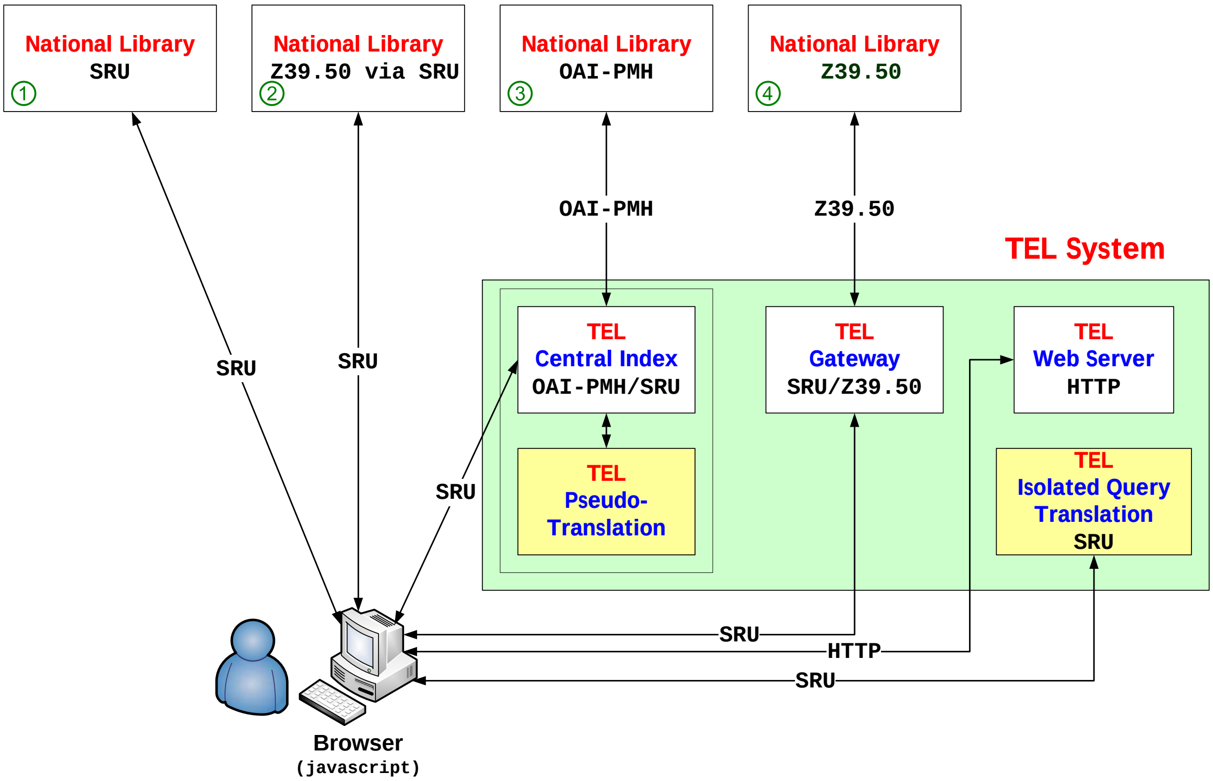


Figure 1: Architecture of the TEL system with the new MLIA functionalities.

Figure 1 shows the architecture of the TEL system with two new components: the first one is going to perform the "isolated query translation", while the second one is going to be responsible for the "pseudo-translation". Note that the "isolated query translation" component can be directly accessed by the client browser by using the SRU protocol and thus the interaction with this new component is actively used by the final user. On the other hand, the "pseudo-translation" component is not directly accessed by the client browser but it represents an extension of the TEL central index, which would be enhanced with MLIA functionalities.

References

[1] Agosti M., Di Nunzio G.M. and Ferro N. 2006. A data curation approach to support in-depth evaluation studies. In F.C. Gey, N. Kando, C. Peters and C.-Y. Lin (eds.), *Proceedings of the International Workshop on New Directions in Multilingual Information Access (MLIA 2006).*

http://ucdata.berkeley.edu:7101/sigir2006-mlia.htm

[2] Braschler M., Di Nunzio G.M., Ferro N. and Peters C. 2004. CLEF 2004: ad hoc track overview and results analysis. In C. Peters, P. Clough, J. Gonzalo, G. J. F. Jones, M. Kluck, and B. Magnini (eds.), *Multilingual Information Access for Text, Speech and Images: Revised Selected Papers from the 5th Workshop of the Cross-Language Evaluation Forum (CLEF 2004) Papers. Bath (UK), 15-17 September 2004*. Lecture Notes in Computer Science 3491. Berlin-Heidelberg: Springer. 10-26.

[3] Braschler M., Ferro N. and Verleyen J.. Implementing MLIA in an existing DL system. In F.C. Gey, N. Kando, C. Peters, and C.-Y. Lin (eds.), *Proceedings of the International Workshop on New Directions in Multilingual Information Access (MLIA 2006).*

http://ucdata.berkeley.edu:7101/sigir2006-mlia.htm

[4] Di Nunzio G.M., Ferro N., Jones G.J.F. and Peters C. 2005. CLEF 2005: ad hoc track overview. In C. Peters, F.C. Gey, J. Gonzalo, G.J.F. Jones, M. Kluck, B. Magnini, H. Müller and M. de Rijke (eds.), *Accessing Multilingual Information Repositories. Revised Selected Papers from the 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005).* Lecture Notes in Computer Science 4022. Berlin-Heidelberg: Springer.

A Study for the Design of Personalization Functions in TEL

Maristella Agosti, University of Padua, agosti@dei.unipd.it
Giorgio Maria Di Nunzio, University of Padua, dinunzio@dei.unipd.it
Yannis Ioannidis, University of Athens, yannis@di.uoa.gr
Julia Luxenburger, Max Planck Institute, julialux@mpi-sb.mpg.de
Gerhard Weikum, Max Planck Institute, weikum@mpi-sb.mpg.de

Background

The European Library[12] is a Web portal which offers access to different resources of 45 national libraries of Europe. This portal has its origin in the TEL project whose aim was to investigate the feasibility of establishing a new service to give access to the combined resources of the national libraries of Europe.

The European Library service is aimed at informed citizens world-wide (both professional and non-professional) who want a powerful and simple way of finding library materials. Moreover, it is expected to attract researchers as there is a vast virtual collection of material from all disciplines. It offers anyone with an interest a simple route to access European cultural resources.

In late 2005 and early 2006, TEL and DELOS[13] jointly review and discuss, from a technological point of view, topics of interest to a future European Digital Library, and to investigate subjects for a more strict cooperation. Among the positive actions that have been jointly undertaken, there is the initiative to study possible new personalization functions to be included in TEL to improve and to make more active the interaction with the final users.

Objective of the Personalization Study

The strategic objective of the study of personalization functions is to produce guidelines and prototype software for new added-value services of interest to the final users, initially selecting those services that present a lower-risk of failure when personalized. A first short-term objective is addressing the development of personalization guidelines, and a second medium-term objective will address integration and testing of existing prototype software and development of possible improvements.

The work is started with an explorative study of the existing access logs from TEL. Main aims of this study are the analysis and categorization of context to derive specifications for new types of logged data and suggestions on the design of innovative personalized services:

- query expansion
- profile building
- notification about new material based on profiles
- recommendations based on profile similarity
- annotation sharing based on profiles
- provision of added-value links and/or service (e.g. OpenURL, etc.), based on preferences and rights of user or organization.

The objective of the TEL log analysis is to provide a personalized access to the visitors of the Web portal. The information gathered from the logs made available by TEL is the basis for the research on creating user profiles (or user community profiles), personalization of search results, etc. The analysis is two-fold: the analysis of the logs of the query issued by users, and the analysis of the http logs of the TEL Web server.

Initial Findings in the Web Server Logs

Ideally, we would like to reconstruct from the HTTP log files a profile for each user that accessed at least once the TEL Website. Therefore, each request should be associated to a unique user identification in order to

[12] http://www.theeuropeanlibrary.org/

[13] http://www.delos.info/

maintain the complete history of the browsing and interaction with TEL Web pages. Some of the ideas for the an ideal personalization of the access to the Web portal are the following:

- The information about the client should be exhaustive enough to localize the user and present the pages with the appropriate language and the appropriate visualization according to the browser adopted.
- The analyses of the server activity should help to understand the different behaviour of those client that highly interact with the Website to those ones that have a quick look to the content of the pages.
- The analyses of the requests should indicate what part of the Website are more visited and what others are not. This information would give directions on how to develop some parts of the Web pages and discard others.

More considerations will arise after the complete analyses of the HTTP log files, that are consistent with the W3C Extended log file format which is the default log file format for Microsoft Internet Information Services (IIS); the adopted version of IIS is the 6.0.

From a deep log analysis, we expect to find interesting relationships among usage and users, and to characterize users by a range of characteristics, including their geographical location, the type of system, the browser used, the frequency of interaction with the Web site, as also suggested by [1]. Different types of techniques to analyze the logs in order to personalize the Web content will be explored. For example, clustering and classification techniques as well as association rules to group users and discover common behaviours [2].

References

[1] Nicholas D., Huntington P. and A. Watkinson. Scholarly journal usage: the results of deep log analysis. In *Journal of Documentation* 61(2): 248-280. http://dx.doi.org/10.1108/00220410510585214

[2] Facca F.M. and Lanzi P.L. 2004. Mining interesting knowledge from weblogs: a survey. In *Data & knowledge Engineering* 53(3): 225-241.

http://dx.doi.org/10.1016/j.datak.2004.08.001

A study on the user interface design for TEL, and on additional user services

Nicola Ferro, University of Padua, ferro@@dei.unipd.it
Norbert Fuhr, University of Duisburg, fuhr@uni-duisburg.de
Emanuele Panizzi, University of Rome “La Sapienza”, panizzi@di.uniroma1.it
Giuseppe Santucci, University of Rome “La Sapienza”, giuseppe.santucci@dis.uniroma1.it
Andreas Rauber, Vienna University of Technology, rauber@ifs.tuwien.ac.at

Background

The European Library (www.theeuropeanlibrary.org) is a Web portal that offers access to different resources of 45 national libraries of Europe. This portal has its origin in the TEL project, whose aim was to investigate the feasibility of establishing a new service to give access to the combined resources of the national libraries of Europe. The European Library service is aimed at informed citizens world wide (both professional and non-professional) who want a powerful and simple way of finding library materials. Moreover, it is expected to attract researchers, as there is a vast virtual collection of material from all disciplines. It offers anyone with an interest a simple route to access European cultural resources.

In late 2005 and early 2006, TEL and DELOS jointly reviewed and discussed, from a technological point of view, topics of interest to a future European Digital Library, and investigated subjects for a more strict cooperation. Among the positive actions that have been jointly undertaken, there is the initiative to study the overall design of the TEL user interface, as well as the exploration of additional services, especially for supporting query formulation, collection navigation and results visualization.

Objectives of the Study

The activities in the short-medium term are focusing on five topics. The main objective is to understand which functions (or services) are more useful in the TEL environment, in order to motivate the end users to explore more large datasets and have “more fun” while exploring TEL. Also expert users like librarians would use some of those services to better visualize (and understand) the existing collections, the results produced by queries and navigation services, and thus be able to define in an easier way customized views for end users.

Evaluation of the present interface. The starting point is a comparative evaluation between the current TEL interface and an appropriate variant of Daffodil (www.daffodil.de), which is an interface to Digital Library functionality developed at the University of Duisburg-Essen. This evaluation will follow both an analytical and an empirical approach. The goal of the analytical evaluation is to assess the functional similarities and differences between the two systems. The goal of the empirical evaluation is to evaluate how well each tool supports the users needs.
The analytical evaluation will consider the usability, functions for search, browsing and result display, and the feedback/help functions of both. For this purpose, the methodology and the questionnaires developed by DELOS will be used. The empirical evaluation will be based on a user-centered and qualitative approach. Its focus is on the users’ experience with the tools, considering user characteristics, preferences and strategies, the types of activities/tasks users perform, and the environment in which the search tool is used. The evaluation must take into consideration the current necessary portal nature of the site.

Support for query formulation. The Daffodil interface already provides functions that help the user in formulating better queries. Most basic, a built-in spell checker will flag search terms not contained in the dictionary, and will propose correct variants. For advanced query formulations, a syntax checker will point out syntactically incorrect formulations. Finally, there is a 'related term' tool that proposes (statistically) similar terms for any of the query terms entered. The comparative evaluation will show to what extent these tools are useful for the TEL users, and then possible integration into the TEL system will be evaluated.

Virtual collections and navigation. It is planned to provide and test an add-on service for building virtual digital collections starting from a set of real ones. For the definition of the virtual collections, the service should distinguish between expert users (like technicians and librarians) and end users. The service relies on automatic batch techniques of indexing, clustering, and classification of existing collections, allowing a visual navigation of their content, for an easier definition of the virtual collections. One of the tools to be tested is SOMLib, developed at the Technical University of Vienna, which is based on the self-organizing map (SOM), a popular unsupervised neural network, used to organize documents by content into topical clusters. Metaphor graphics facilitate the intuitive representation of the resulting document archive, allowing users to get an instant overview of its content and characteristics. The benefits expected are that the users can be presented with cross collection views, can deal with smaller set of more relevant data and therefore queries can be processed in a faster way.

Presentation/visualization of results. The Daffodil system already provides functions for relevance ranking or quick filtering of results, as well as extracting attributes like author names or frequent terms from the result set. In addition, it is planned to provide and test an add-on service that allows end users to interact with the query result in a more effective way. Several techniques can be used in this context: real time indexing, cluster-gather algorithms, smart use of relevance factor, information visualization techniques. One of the tools to be tested is DARE (Drawing Adequate Representation, developed at the University of Roma1), which is a visualization tool allowing the user to analyze large amounts of data. An innovative feature of the system is that it supports at the same time disaggregate and aggregate (OLAP) data visualizations, allowing the user to move seamlessly between them. Moreover, the system incorporates a knowledge base that automatically produces the best visual data representation for a given data set, avoiding the end user to deal with this time consuming and error prune activity.

Annotation. Annotations support the idea of collaborative search and enrichment of the Digital Library, as more and more users want to participate and share information, acting also as contributors rather than just as simple readers. The three main functions to support this notion are: (i) adding information related to documents or to query results; (ii) categorizing, cataloguing and linking resources; (iii) adding new content and discussing others' contributions. This activity will be based on two tools (MADCOW, developed at the University of Roma1, and FAST, developed at the University of Padua). In the first step there will be the possibility to annotate one record or a result set, and to show all annotations available for that record or for a result set. In a second step it will be possible to make queries also on the contents of the annotations and to annotate different parts of records (tags, multimedia objects, upload/attach media objects).

Printed by ISTI-CNR
At GRUPPO ALI, Pisa, Italy
August 2006

ISSN 1818-8044
ISBN 2 -912335-22-1