Dagstuhl Seminar 99351
August 29 - September 3 1999

Multimedia Database Support for Digital Libraries

E. Bertino (Milano)
bertino@dsi.unimi.it

A. Heuer (Rostock)
heuer@informatik.uni-rostock.de

T. Ozsu (Alberta)
ozsu@cs.ualberta.ca

G. Saake (Magdeburg)
saake@iti.cs.uni-magdeburg.de

Contents

1 Motivation

Digital libraries are a key technology of the coming years allowing the effective use of the Internet for research and personal information. National initiatives for digital libraries have been started in several countries, including the DLI initiative in USA http://www-sal.cs.uiuc.edu/ sharad/cs491/dli.html, Global Info http://www.global-info.org/index.html.en in Germany, and comparable activities in Japan and other European countries.

A digital library allows the access to huge amounts of documents, where documents themselves have a considerably large size. This requires the use of advanced database technology for building a digital library. Besides text documents, a digital library contains multimedia documents of several kinds, for example audio documents, video sequences, digital maps and animations. All these document classes may have specific retrieval methods, storage requirements and Quality of Service parameters for using them in a digital library.

The topic of the seminar is the support of such multimedia digital libraries by database technology. This support includes object database technology for managing document structure, imprecise query technologies for example based on fuzzy logic, integration of information retrieval in database management, object-relational databases with multimedia extensions, meta data management, and distributed storage. The seminar is intended to bring together researchers from different areas like object and multimedia databases, information retrieval, distributed systems, and digital libraries. It is the intention of this seminar to clarify differences in terminology between these areas, to analyze the state of the art, discuss requirements of digital libraries for multimedia databases and to identify future trends in research and development. The seminar should therefore focus on two major questions:

- Which functions of digital libraries need database support?
- What can database techniques offer to support these digital library functions?

These major questions can be detailed to specific topics, which list the technological areas relevant for this seminar (this list is of course not exhaustive):

- How to support digital libraries?
  - Document Servers
  - Supporting Different Types of Documents in Database Systems
  - Document Acquisition and Interchange
  - Extending Object-Relational and Object-Oriented Database Technology for Digital Libraries
  - Storing, Indexing, and Querying Large Sets of Documents
  - Integration of Heterogeneous Meta data and Documents
  - Combining Querying on Structured Meta data and Content-based Retrieval
  - Integrating Vague and Fuzzy Queries
  - Distribution of Queries

  - Different User Views on Large Sets of Documents
  - Visual Interfaces to Digital Libraries

- Which features of digital libraries need database support?
  - Alerting Services
  - Intelligent User Agents, Personal Digital Libraries
  - High Performance Document Servers
  - Efficient Retrieval Functionality
  - Document Delivery and Data Dissemination
  - Security and User Access Models
  - Trusted Document Servers

2 Agenda

Monday Morning

Opening

Actual Systems

A. Desai Narasimhalu, Kent Ridge Digital Labs. - Singapore
Digital Library Efforts in Singapore

Gerhard Möller, Universität Oldenburg
Gerhard: Navigating the Web with the Universal Decimal Classification System

Michael Ley, Universität Trier
DBLP

Susanne Boll, Universität Ulm
Gallery of Cardiac Surgery - Database Support for a Multimedia Repository

Michaela-Monica Vladoiu, Petroleum-Gas Univ. - Ploisti
Multimedia Databases and Computer On-line Learning

Monday Afternoon

Queries & Interfaces I

Kasim Selcuk Candan, ASU - Tempe
Query Optimization in Multimedia Databases

Ruxandra Domenig, Universität Zürich
The Singapore Approach for Querying Heterogenous Data Sources

Iztok Savnik, Universität Freiburg
Algebra for Distributed Data Sources

Tuesday Morning

XML and SGML Databases

M. Tamer Ozsu, University of Alberta
SGML Databases

Holger Meyer, Universität Rostock
Managing and Querying XML-Documents with Object-Relational DBMS

Jürgen Wäsch, GMD-IPSI, Darmstadt
XML and the Evolving Digital Economy

Klemens Böhm, ETH Zürich
Efficient Semistructured Data Management in Power DB

Tuesday Afternoon

Queries & Interfaces II

Ron Sacks-Davis, RMIT - Carlton
Architectures for Structured Document Data

Arjen de Vries, University of Twente
Content Management and Database Query Processing

Isabel Cruz, WPI - Worchester
A User Interface for Distributed Multimedia Database Clustering with Mediator Refinement

Tiziana Catarci, Universita di Roma "La Sapienza"
Laurin, A European Project to Build a Clipping Digital Library

Tuesday Evening

Ralf Nikolai, FZI Karlsruhe
Semantic Integration by the Integration of Retrieval Vocabulary

Wednesday Morning

System Issues

Ulrich Marder, Universität Kaiserslautern
Improving the Performance of Media Servers Providing Physical Data Independence - Problems, Concepts, and Challenges

Henrike Berthold, Klaus Meyer-Wegener, TU Dresden
Interoperability of Media Servers and OO/ORDBMS

Wolf-Tilo Balke, Werner Kießling, Universität Augsburg
The HERON Project: Experiences in Using Object-Relational Databases for Content-based Image Retrieval in Digital Libraries

Annika Hinze, FU Berlin
Alerting Services for Digital Libraries

Thursday

Queries & Interfaces III

Norbert Fuhr, Universität Dortmund
Logic-based Retrieval in Multimedia Digital Libraries

Ingo Schmitt, Universität Magdeburg
A Comparison of Multimedia Query Languages

Discussion Groups

- **Datamodels for Digital Libraries**
- **Information Quality**

Discussion Groups

- **Query Languages for Digital Libraries**
- **Business Models for Information Interchange**

Discussion Groups

- **The Web and its Impact on Architectural Issues of Digital Libraries**
- **Special Applications of Digital Libraries**

Friday

Reports from the discussion groups

Closing

3 Abstracts

Digital Library Efforts in Singapore

A. Desai Narasimhalu, Kent Ridge Digital Labs. - Singapore
arcotdesai@yahoo.com

Digital Libraries are defined to be digital content that are managed using IT resources. Singapore has a nationwide backbone nework called the Singapore ONE that has 2 Mbps downstream and 200 KBps upstream capacity, that allows accessing digital library from homes. Singapore has also established National Library Board and set aside one billion Singapore dollars to develop digital libraries while retaining access to legacy collection. KRDL has developed technologies such as multilingual information retrieval, federated database systems, machine translation, video streaming, Asian Language processing which were integrated into Phase 1 of the digital library implementation. The current digital library allows for access to bibliographic information using TIARA application, business information using electronic databases, VEGAS application for image and text collections, CD-ROM and video on demand. Singapore has selected on educational and business sectors as their major areas of focus. Singapore is also developing an A5 size portable device to address the last mile problem - allowing users to be mobile within a limited distance while remaining connected to a digital library. Phase 2 efforts will extend the basic services to all the households (70aspects of digital library.

Gerhard: Navigating the Web with the Universal Decimal Classification System

Gerhard Möller, Universität Oldenburg
Gerhard.Moeller@OFFIS.Uni-Oldenburg.de

GERHARD (German Harvest Automated Retrieval and Directory, http://www.gerhard.de/) is a fully automatic indexing and classification system of the German World-Wide Web for integrated searching and browsing. A database-driven robot collects academically relevant documents, which are automatically classified with computer-linguistic and statistical methods using the Universal Decimal Classification. The generated metadata and the index of the documents are held in a relational database (Oracle with Context option). The user-interface is trilingual (German, English, French) and allows the user to look for "similar" documents very easily through its tight integration of searching and browsing mechanisms.

The talk first showed some typical problems with IR on the Web and offered a possible solution (GERHARD). After the system was presented, I shared the lessons we learned with the audience to finally draw the conclusion on how the digital library community could benefit from the results of GERHARD:

- the use of the UDC can be encouraged,

- the automatic classification actually performs better than one could expect and can even be improved for DL, as the documents in DL are more homogenous and speed does not matter that much,
- integration of searching and browsing is necessary for non-expert users to exploit the complex UDC, and
- multimedia-support is an open problem that might be tackled by exploring link-information.

DBLP

Michael Ley, Universität Trier
ley@uni-trier.de

DBLP is a free Computer Science bibliography available on the Web. It provides basic bibliographic information for more than 129000 papers published in journals and proceedings. Visit http://dblp.uni-trier.de.

The ACM SIGMOD Anthology is a CDROM publication by the ACM Special Interest Group on Management of Data which contains full texts (PDF) from SIGMOD Conf. (1975-1997), the Very Large Data Bases conference (1975-1997), the Symposium on Principles of Database Systems (PODS, 1982-1998) and the IEEE Data Engineering Bulletin (1993-1998). These full texts are combined with DBLP to form a digital library for the database systems research field. An extended version with TODS and the Data Engineering Conference is under construction. The talk gives a short overview about the production process of DBLP and the Anthology.

Gallery of Cardiac Surgery - Database Support for a Multimedia Repository

Susanne Boll, Universität Ulm
boll@informatik.uni-ulm.de

The project "Gallery of Cardiac Surgery" aims at developing an Internet-based, database-driven multimedia information system for physicians, medical lecturers, students, and physicians in the domain of cardia surgery.

The users are provided with multimedia information according to their user specific request to the multimedia information system, their background knowledge, their different understanding of the subject, their location and technical infrastructure. Global requirements to the design of the system are the extensive reuse of the expensively produced multimedia content by different users in their speific context, the user context specific delivery of multimedia content, and presentation-neutrality, i.e., the separation of layout and structure to separate storage from delivery and presentation.

The databases and information systems group (DBIS) in Ulm is developing the database driven multimedia repository. The talk presented the approaches in modeling media data, metadata, and multimedia documents to give suitable database support for the

management of the multimedia data, content-based retrieval, authoring of multimedia content, and the delivery and presentation of multimedia documents. The approach taken is to extend object-relational database system technology by the needed data types and functionality. The talk presented the extensions (DataBlades) for the object-relational Informix Dynamic Server/Universal Data Option the group developed so far: the media integration blade and the ZYX blade. Additionally, the talk showed how this database support can now be employed and exploited by the integrated tools we develped for managing, retrieving, annotating, streaming, delivering, and presening the content, and by thi to realized the desired featrues of the multimedia repository

Trends and future demands: most economical (re-)usage, individualization of content, flexible distribution and delivery of content.

Multimedia Databases and Computer On-line Learning

Michaela-Monica Vladoiu, Petroleum-Gas Univ. - Ploisti
monica@csd.univ.ploiesti.ro

We may say the knowledge is the new global asset - so people have to be continually learning, crafting innovative solutions to changing circumstances, staying informed and responsive. Computer Based Learning (CBL) provides perhaps the best opportunity for person self-guided learning. If we take into consideration the place of Internet in everybody's life nowadays and if we furthe make a step forward from CBL at a learning process using the net, we reach a more powerful variant of CBL: Computer On-line Learning (COL). If we keep in mind these two important needs we can easily see the general and efficient organization of MM information is one of the main requirements for developing courseware which "can get into" anyone's computer quickly and "softly". Multimedia DataBase Management Systems (MM DBMSs) offer a good solution for this problem. They allow to store and manipulate such data in an efficient manner with a high level of generality. Vast digital libraries of information will soon be available on the Internet as a result of emerging technologies for manipulating multimedia. These libraries will profoundly impact the educational activities.

Query Optimization in Multimedia Databases

Kasim Selcuk Candan, ASU - Tempe
candan@asu.edu

Abstract: In this presentation, I provided an overview of the challenges associated with the processing of multimedia database queries. I used the Semantic and Cognition-based media retrieval (SEMCOG), Collaborative Heterogeneous Interactive Multimedia Platform (CHIMP), and Virtual Reality Repository (Vrep) projects as the basis of the observations.

Multimedia databases have to deal with the semantic (spatio-temporal dimensions, user- and context-dependence, subjectivity, and availability at various quality levels) as

well as physical (volume, quality/cost trade-off, and interactivity) heterogeneities, simultaneously. Consequently, a multimedia database must provide transparency against inherent heterogeneity and inherent fuzziness. In this presentation I talked about different sources of fuzziness in multimedia data (including similarity/correlation in media features, imperfections in the feature extraction process, imperfections in the query formulation, imperfections in the available index structures, and partial match requirements. I highlighted the needs for the development of meaningful merge functions and progressive algorithms that can deal with the fuzzy nature of multimedia data. Since the solution space is generally very large, these algorithms should be able to provide approximate results, using statistics and known properties of predicates and merging functions. Furthermore, since a multimedia query may return a different number of results depending on the execution plan, and since a predicate may have different "qualities" depending on the execution/binding patterns, query optimization must not only be aimed at getting the cheapest execution pattern, but also the highest expected quality. In this talk I have also highlighted the need for using "structural information" (such as time, space, interaction, and hierarchy) in answering multimedia (or Web) related queries. This task however requires metrics for comparing various structural models and algorithms that can progressively return results based on these metrics. Such structural information can be used both in indexing structured data as well as in summarizing query results to match users' needs and resources. Finally, I have pointed to the need for feedback (hints) provided by the system to guide and inform the user in query formulation and reformulation process. To achieve this goal, the database system must harness statistics that will act as feedback to the users. Furthermore, query results must be visualized as "presentations" that users can browse through, while the system gathers user/query profiles to prefetch objects that a user may intend to view in a near future.

I have concluded the talk by reminding that designing a system for multimedia object/document management is challenging, with many novel research issues waiting to be solved; and that we have some pieces, but pieces are not enough yet.

The Singapore Approach for Querying Heterogenous Data Sources

Ruxandra Domenig, Universität Zürich
domenig@ifi.unizh.ch

Large amounts of data spread over heterogeneous data sources like traditional database systems, information retrieval systems, data sources managing semistructured data are accessible on-line today. However, the "technical" availability alone is not at all sufficient for making meaningful use of the existing information. Therefore, the problem of effectively and efficiently accessing and querying heterogeneous data sources is an important research direction. SINGAPORE (SINGle Access POint for heterogeneous data REpositories) is a system which aims at retrieving information from a network of heterogeneous data sources. We discuss the requirements and architecture of SINGAPORE and then present the query language, which allows to query structured, semistructured and unstructured data sources in a unified way. We finally discuss the special aspects of query processing in such a heterogeneous environment.

Algebra for Distributed Data Sources

Iztok Savnik, Universität Freiburg
savnik@informatik.uni-freiburg.de

The Internet contains large amount of different data sources accessible through ftp files, XML or HTML documents, and various types of wrappers around relational and object-relational database systems. The data available via such data sources may vary from simple lists of records, catalogs containing large amounts of flat tables, to complex data repositories including large conceptual schemata and tables composed of complex objects. The data about the particular subject of interest can be spread among the number of distributed sites among which the communication speed can vary significantly.

Querying distributed data source poses several new problems such as, for example, uniform access to the data sources, integration of the information provided by the data sources, and efficient manipulation of large amounts of data in such environments. Our work in this area focuses on the design and implementation of the query execution system for the distributed data sources. In particular, we investigate the algebra which allows for the efficient manipulation of data from the distributed data sources, and the implementation of the query optimizer which minimizes the data transfer and the computation time involved in processing of queries over distributed data sources.

The design of the algebra for the distributed database sources relays on the existing work on the relational and object-relational query execution systems. The kernel of the algebra comprises relational operations extended for the manipulation of complex objects. The new constructs are included in the algebra to reflect the specific properties of the data environment. Firstly, the algebra is based on the formal view of object-oriented and frame-based languages provided by F-Logic. The data model provides a convenient environment for the representation of semi-structured as well as structured data. Further, the algebra includes a set of operations for querying conceptual schemata. These operations allow for browsing the conceptual representation of the data sources and using the elements of the conceptual schemata for querying extensional data. Finally, the algebra includes the operations for the manipulation of data distributed among several sites. The most important operation in this respect, which was originally suggested by Freytag in [Freytag:SIGMOD87], is the operation for shipping objects among the sites of the network. Its integration in the algebra allows uniform approach to the optimization of the data transfer among the sites and the time required for processing of local queries.

SGML Databases

M. Tamer Ozsu, University of Alberta
ozsu@cs.ualberta.ca

Digital libraries require the storage of multimedia documents. We describe the design of an object-oriented multimedia database management system that can store and manage SGML/HYTIME compliant multimedia documents. The system is capable of storing, within one database, different types of documents by accommodating multiple document

type definitions (DTDs). This is accomplished by dynamically creating object types according to element definitions in each DTD. The system also has tools to automatically insert marked-up documents into the database. We discuss the system architecture, design issues and the system features.

The talk is based on the following paper: M. T. Özsu P. Iglinski, D. Szafron, S. El-Medani, M. Junghanns. "An Object-Oriented SGML/HYTIME Compliant Multimedia Database Management System", Fifth ACM International Multimedia Conference (ACM Multimedia '97), Seattle, WA, November 1997, pages 239-249.

Managing and Querying XML-Documents with Object-Relational DBMS

Holger Meyer, Universität Rostock
hme@informatik.uni-rostock.de

Nowadays, a lot of applications around the web use or intent to use XML as an intermediate format for representing large amounts of structured or semi-structured data. Relational and object-relational database systems are a well understood technique for managing and querying large sets of structured data.

In our approach the nested relations data model is the basic model for representing XML data in object-relational database systems. Using the Partitioned Normal Form (PNF) we show how a relevant subset of XML documents and their implied structure can be mapped to database structures. Beside straight forward mappings, there are some XML structures which cannot be easily mapped to database structure, namely mixed content and alternatives. These structures would result in large database schemas and sparse populated databases. As a consequence, such XML document fragments should be mapped to database attributes of type XML and kept as is.

We present an algorithm which finds some kind of optimal mapping based on the XML Document Type Definition (DTD) and statistics. The statistics are derived from sample XML document sets and some knowledge about queries on XML document collections.

XML and the Evolving Digital Economy

Jürgen Wäsch, GMD-IPSI, Darmstadt
waesch@darmstadt.gmd.de

Information Commerce is an evolving area in electronic business. Intermediate information brokers are collecting information from various distributed, heterogenous, and autonomous sources (e.g., WWW-Sites) and remanufacture them into new products. This process includes search and retrieval of relevant content, combination and enrichment of information as well as the individualization and dissemination of value-added services and information products.

As an exemplary scenario we constructed a brokering service for golf players that help them to choose a course. It integrates three independent information sources: a guide of golf sites, a route planner, and localized weather forecast. A typical query would ask for

all 18-hole golf courses up to 200 km from Dagstuhl, where the rain probability is less than 10 percent and the green fee is below 50 DM.

In this talk, we present the overall architecture of GMD-IPSI's XML information brokering framework. It utilizes the eXtensible Markup Language (XML) as a common exchange and data format to overcome the syntactic heterogeneity of different sources. The framework consists of several tools which enable developers to rapidly build integrated information services. For information extraction and XML generation we offer the JEDI toolkit that includes generic wrappers as well as a failure-tolerant parser to extract data from irregularly structured documents. The extracted information is stored and administered in an XML warehouse that builds upon our persistent implementation of W3C's Document Object Model (PDOM). Declarative queries can then be used to form integrated views on the information modeled in the warehouse. For this, the framework offers a query processor that builds upon the XQL proposal of the W3C.

One important issue in such an information brokering architecture is efficient storage of XML documents in the middle tier. To that end, we discuss the state of the art in the area of XML stores and XML-enabled database systems. We present different commercial products and prototypes and discuss their pros and cons. Other important questions that are raised in this talk are the need for a commonly agreed data model for XML and for standard query and schema languages for large collections of XML documents. Such languages are currently under development and will contribute to the solution of the semantic information integration problem which is currently not addressed by the presented brokering framework.

Architectures for Structured Document Data

Ron Sacks-Davis, RMIT - Carlton

Semi-structured data, including but not limited to structured documents, has specific characteristics and is used in many ways different to tabular data. SGML and XML are widely used to represent information of this type. The demands on systems that manage semi-structured data vary from those on traditional relational systems. This talk reviews the nature and characteristics of semi-structured data, and the functional needs of those applications, including query requirements, document description, manipulation, and document management needs. We examine alternative physical models for semi-structured data and evaluate and compare alternative system architectures.

We describe a content management system, called the Structured Information Manager (SIM) which provides native support for XML and SGML. In SIM, the DTDs are directly incorporated into the database schema and the SGML/XML documents are stored in their native format.

Content Management and Database Query Processing

Arjen de Vries, University of Twente
arjen@cs.utwente.nl

The current research in "multimedia databases" has resulted in ADT support for multimedia retrieval. We argue that this does not offer sufficient support, and suggest what features are lacking: querying should be iterative, and use multiple representations, while the user only provides examples and relevance judgements. Furthermore, a multimedia DBMS should provide content independence, with which we mean that the retrieval engine is independent from the currently available metadata extraction techniques.

The proposed solution, the "Mirror DBMS", integrates IR techniques into the DBMS, using an adapted version of the inference network retrieval model. The strict separation between the logical level and the physical level provides better opportunities to optimize query processing. The resulting system has been implemented using Moa object algebra at the logical level, Monet at the physical level, and special Moa extensions for IR processing. The prototype is demoed with a small image collection, using two color and four texture spaces, in combination with manual annotations. The Mirror DBMS is also participating in TREC-8. The complete integration enables the user to further constrain the content search by traditional data retrieval.

A User Interface for Distributed Multimedia Database Clustering with Mediator Refinement

Isabel Cruz, WPI - Worchester
cruz@cs.wpi.edu

The DelaunayMM system supports an interactive, customizable interface for querying multimedia distributed databases, like Digital Libraries. Through this interface, users select virtual document styles that cater the display of query results to their needs, while also offering transparent pre- and post-query refinement and nested querying. DelaunayMM's virtual documents preserve context by maintaining a single customizable interface for result viewing. The advanced transparent query features rely on mediation to provide adept access to information. In this paper, we present the framework for DelaunayMM, its architecture, the user interface, and results of the first usability study.

Laurin, A European Project to Build a Clipping Digital Library

Tiziana Catarci, Universita di Roma "La Sapienza"
catarci@infokit.dis.uniroma1.it

The field of digital libraries has been attracting a lot of research efforts during the last years. Many interesting projects have been started, dealing with the various open issues arising in the field. However, no project has specifically taken into account the problem of building a digital library of newspaper clippings. It is well known that a huge part of

cultural knowledge is stored in the newspapers of yesterday. Since newspapers were not always easily accessible, special clipping archives have been created in the 20th century. People interested in newspaper information benefit from these archives because the work of selecting, cutting and indexing articles is done by specialists. In order to maintain their important position in the information market, clipping archives should be able to integrate their special skills (such as professional knowledge and experience in gathering and treating newspaper information) into the new technologies of the information society. The EU-funded Laurin project will carry out the preliminary work necessary for an efficient and smooth shift from the "analogue" clipping archive to its "digital" successor. In order to effectively accomplish this hard task, the Laurin Consortium has gathered a significant number of libraries, which are acting as final users and test sites and are continuously driving the system design and development with requirements, suggestions, testings, and criticisms. The talk presented the Laurin design methodology, the main user and organizational requirements for a clipping digital library, and the overall architecture of the Laurin system.

Semantic Integration by the Integration of Retrieval Vocabulary

Ralf Nikolai, FZI Karlsruhe
nikolai@fzi.de

Semantic integration is of major importance, e.g., if distributed information systems are integrated to provide a common entry point not only on a technical, but also on a semantic level. An important step towards a semantic integration for search and retrieval purposes is the integration of retrieval vocabularies.

Thesauri are widely used to provide such a retrieval vocabulary. Therefore, our research concentrates on the process of loosely interconnecting the vocabularies of different thesauri by establishing interthesaurus relationships and handling inconsistencies. In contrast to other approaches, our approach for thesaurus integration carefully takes into account the results of a preceding compatibility evaluation. A catalogue of evaluation criteria is presented. Practical experiences we gained in applying these criteria to two real-world thesauri from related domains (general environment, agriculture) identified significant problems, which are inherent to the integration of classicalthesauri according to ISO norms. Based on these experiences we have developed an approach to enrich the information structure of classical thesauri. Additionally, we introduce a 6 phase integration model based on this enriched information structure.

Improving the Performance of Media Servers Providing Physical Data Independence - Problems, Concepts, and Challenges

Ulrich Marder, Universität Kaiserslautern
marder@informatik.uni-kl.de

This presentation deals with media servers providing physical data independence. Todays media serversespecially continuous media serversusually do not provide physical data

independence at all. Oneif not the mainreason for this is performance. Physical data independence without optimization costs a lot of performance. Therefore, we are looking for a solution of this optimization problem.

The following is a typical scenario where physical data independence would be highly beneficial: There is global media dataoften called media assetsstored in an MM-DBMS (which might be a part of a digital library). Lots of heterogeneous clients with different capabilities of storing, processing, and presenting media data are willing to access this MM-DBMS. Some only want to retrieve media objects for presentation or possibly printing. Others create or modify media objects. And again others create media objects by editing and combining existing ones. Assuming a sort of unbalance between these applications sounds reasonable: Usually there will be many applications of the presentation type but few of the other types. Thus, one could think of optimizing the system for presentation. But unfortunately, the other applications often have much stronger quality and performance demands.

We show that there are considerable problems when attempting to provide physical data independence with a media server (or MM-DBMS). Such systems tend to require frequent format conversions inevitably resulting in bad performance. They may inadvertently lose data due to irreversible updates. And hiding the internal data representation from the client obviously means that all the strongly necessary optimization is to be accomplished by the server, which is both more difficult and more promising than leaving optimization to the applications.

Our proposed media server concept is based on a generalization of data independence called transformation independence. This abstraction reduces the creation, retrieval, and modification of media objects to what can be called the "pure application semantics". The consequence are multiple optimization dimensions being left for exploitation by the server. The VirtualMedia concept realizes transformation independence based on virtual media objects being described by filter graphs. With this concept, optimization can be basically characterized as the process of optimally matching transformation request graphs and materialization graphs.

Since we are just at the beginning of developing these concepts, there are still a lot of challenges to be mastered, like formalizing and evaluating the graph transformation algorithms and realizing the concept using available DBMS-technology and media server components.

Interoperability of Media Servers and OO/ORDBMS

Henrike Berthold, Klaus Meyer-Wegener, TU Dresden
berthold@db.inf.tu-dresden.de

Information systems in general manage structured data, and most of them use databases to store them adequately. While these systems work well, there are new requirements now to improve them towards the inclusion of multimedia data, i.e. images, graphics, video, and audio. Usually, multimedia data are stored in specialized servers which can cope with the requirements of real-time storage and delivery. Applications being developed today, however, need the services of both databases and these media servers. We propose a Federated Multimedia Database System that has three types of data sources, namely

media servers, traditional databases and multimedia retrieval engines. In this talk we discuss the integration aspect, especially the design of the global object-oriented database schema and the processing of OQL queries using the intermediate representation called monad comprehensions.

The HERON Project: Experiences in Using Object-Relational Databases for Content-based Image Retrieval in Digital Libraries

Wolf-Tilo Balke, Universität Augsburg
Balke@Informatik.Uni-Augsburg.DE

Today multimedia database systems are a most challenging area of database research. Traditional text-based information systems are increasingly extended to integrate multimedia data as for instance images, audio and video files. One special group of these extended systems are image databases for storage, management and retrieval of still images. At present, the most common technique for integrating images into a database is to store them together with some descriptive text or keywords assigned by human operators and retrieve them by matching the query texts with the stored keywords. Those texts assigned are not only very subjective and incomplete, but also very expensive. A far more promising technique -known as content-based retrieval- captures the image content automatically by visual features as colors, textures or the shapes of image objects.

The interdisciplinary HERON-project investigates the impact of multimedia database technology and content-based retrieval on broad application areas in the humanities. As a first area of application heraldry has been chosen due to the large heraldic image archives. The HERON-project concentrates on merging application semantics and image processing techniques to gain adequate features for content- based retrieval, thus optimizing the relevance of result sets. A very important preprocessing step is the automatic segmentation of images for shape retrieval.

As image queries in general consist of multiple features combining visual features with fulltext retrieval, another main area of research in the scope of the HERON-project are efficient combination of ranked result sets and multimedia delivery. Therefore a middleware has been proposed splitting complex image queries according to their atomic features, collecting and efficiently combining the atomic ranked result sets. Finally the overall best results in a large variety of image formats and qualities are delivered according to the specific user's needs. Using a so called format optimization the set of physically stored and at delivery time converted image formats and qualities is chosen due to specific characteristics of both the image server and the application profile. This set can be automatically updated as the application profile changes.

Alerting Services for Digital Libraries

Annika Hinze, FU Berlin
hinze@inf.fu-berlin.de

In the last years, alerting systems have gained strengthened attention. Several systems have been implemented. For the evaluation and cooperation of these systems, the following problems arise: The systems and their models are not compatible, and existing models are only appropriate for a subset of conceivable application domains. Due to modeling differences, a simple integration of different alerting systems is impossible. What is needed, is a unified model that covers the whole variety of alerting service applications.

In this talk I introduce a unified model for alerting services that captures the special constraints of most application areas. The model can serve as a basis for an evaluation of alerting service implementations.

In addition to the unified model, we define a general profile structure by which clients can specify their interest. This structure is independent of underlying profile definition languages.

To eliminate drawbacks of the existing non-cooperating solitary services we introduce a new technique, the Mediating Alerting Service (MediAS). It establishes the cooperation of alerting services in an hierarchical and parallel way.

Logic-based Retrieval in Multimedia Digital Libraries

Norbert Fuhr, Universität Dortmund
fuhr@ls6.informatik.uni-dortmund.de

We discuss five concepts which are relevant for content-based access of mutimedia data: 1) Uncertain inference is required due to the vagueness of information needs and the imprecise representation of the semantics of multimedia objects. 2) Whereas text retrieval is based on proposition loggic, some kind of predicate logic is required for dealing e.g. with spatial or temporal relationships in multimedia. 3) In order to select relevant subtrees of hierarchically structured documents, we propose the concept of augmentation. 4) Retrieval of content should be based on an open world assumption, whereas retrieval of facts should use the closed-world assumption. 5) Possible inconsistencies in the data should be handled by an appopriate inference mechanism, e.g. a four-valued logic. As a prototype implementation of these concepts, we present the DOLORES system, which is based on a probabilistic, object-oriented four-valued logic.

A Comparison of Multimedia Query Languages

Ingo Schmitt, Universität Magdeburg
schmitt@iti.cs.uni-magdeburg.de

Query languages are playing a very important role as interface between users and database systems. Well-known query languages like SQL or OQL are designed for relational and object-oriented databases, respectively, but are not suited for multimedia databases. In the literature there are many different proposals of new multimedia query languages. The presentation compared the proposed languages in order to find out general and specific deficiencies which can be the basis for further developments. The underlying comparison criteria focus especially on the data model concepts like content-based retrieval, spatial and temporal relations, and presentation of multimedia objects. Basing on the comparison results different classifications of the proposed languages have been proposed and needs for further research were presented.

4 Discussion Group Summaries

Data Models for Digital Libraries

Prepared by Klaus Dittrich and Jürgen Wäsch

The focus of this working group was to discuss the (rather general) issue of data models for digital libraries.

Since the notion of a digital library is very broad and a deep terminological discussion was outside the scope of this group, we first came up with a "working definition" of a digital library. We characterized a digital library as a combination of (1) digital content (i.e., documents, content-related meta-data, etc.) with (2) computerized services (indices, catalogs, functions on the digital content, etc.). Thus, besides the content itself, a digital library has to offer services that are comparable to a human librarian in a traditional library (to whatever extent).

Afterwards, we discussed what the purpose of a data model in the context of digital libraries should be. Answers were, e.g., description of structure, content, semantics, and layout of documents, relationships among documents, document retrieval and query processing/optimization, storage and updates of documents, exchange of contents among digital libraries, and integration of content and services in federated digital libraries.

A question that came up was where, i.e., at which abstraction level, data models apply in the context of digital libraries (if they matter at all). In order to systematize the discussion, we used the *three layer architecture* of database management systems: (1) physical layer, (2) logical layer (here MMDBMS technology applies) and (3) conceptual layer (here digital library concepts apply).

- *Logical layer*: Data models belonging to the logical level can be categorized as follows: (1) DB-based models (relational, object-oriented, object-relational), (2) IR-based document models (models underlying SGML, XML, etc.), (3) semi-structured data models like OEM, and (4) knowledge representation.

- *Conceptual layer*: We discussed which features a conceptual model for digital library should provide, i.e., the universe of discourse in digital libraries. A number of diverse features were mentioned: fuzzyness, reasoning, abstractions, semantic descriptions, coordinated streams of continuous media, multiple representations, copyright, client models, and business models.

  It was felt that multiple representations are important in digital libraries, since we can not assume a single content format. Moreover, conceptual models are likely to be application-dependent. This raised another issue: coexistence and integration of several conceptual models which is especially important in the area of federated digital libraries.

Another important issue is the mapping of the conceptual model(s) to the data model(s) provided at the logical layer. The question here is whether the data models are sufficient to express the features of the conceptual model and whether information is lost in the mapping from the conceptual layer to the logical layer. In general, it was felt that

the data models at the logical layer must be extensible in order to adequately represent the diverse features of the conceptual layer. How much the system must know about data model extensions was left as an open issue.

In the following, we summarize other important issues that were discussed throughout the session.

Yet another data model? A question was if we should start from scratch or reuse a data model/conceptual model and extend it towards digital libraries. Some people were sceptical about new proposals. They argued that it is easy to come up with a new model, especially in a research environment, whereas it is difficult to implement the model. So, maybe a better question is: what is the best extension of a given model with respect to digital libraries?

XML as a data model for digital libraries. An issue that was brought up was whether XML can serve as a data model for digital libraries. The common feeling was that digital library content is much more complex than XML documents. It seems to be difficult to map the higher level concepts to XML without losing information, since XML has limited expressiveness, e.g., no data types or constraints. Moreover, XML itself has no semantics assigned—it is merely syntax (in contrast to RDF and other domain-specific applications and vocabularies build upon the XML syntax).

Do standards matter for digital libraries? The opinion was that they definitely matter at some point, since large digital libraries will be federated systems (consisting of specialized storage systems and digital libraries themselves).

Data models vs. business models. A common consensus was that, besides data models, business models are important in the context of digital libraries. Concepts like ownership, copyrights, usage rights (e.g., rights to disseminate, borrow, sell (parts of) the information) play an important role, as well as payment models (e.g., pay per view, pay per use). Business models for digital libraries were further discussed in a separate discussion group.

What data model do you prefer? The discussion session was closed by asking the participants which data model and technology they would apply if they should built a digital library within a two years timeframe. Among the answers were: (1) object-relational database technology together with XML, (2) file system together with external indices and the relational model for meta-data, and (3) specialized storage servers with an integration layer on top (problem: close integration of multimedia objects with respect to their structure).

Query Languages for Digital Libraries

Prepared by K. Selcuk Candan

We started with discussing what type of a query language would be appropriate for digital libraries. One suggestion was to divide the current languages into four categories: (1) general purpose languages (such as SQL99, OQL, logic-based languages), (2) structure oriented languages (such as XQL, XMLQL, LOREL), (3) media oriented languages (such as SQL/MM, MOQL, CSQL/VCSQL), and (4) information retrieval oriented languages (such as Z.39.50). We also identified the reason why SQL was (arguably) a success as the fact that it satisfied the needs of the target applications, i.e., financial institutions. Hence,

we tried to identify the needs of the target DL applications. This line of thought lead to the observation that we need to differentiate between the query languages for end users (LUs) and the query languages for developers (LDs). The general agreement was that database researchers need to be involved in (or at least be aware of the issues involved in) the design of end-user languages, as databases may need to provide certain functionalities that LU developers need.

The discussion group observed that a query language for end users (LU) should (1) be intuitive/easy to use, (2) address the (potentially conflicting) needs of experts and non-experts users, (3) be descriptive, and (4) be incremental. This last observation is due to the fact that human search process itself is mostly characterized as incremental. In order to understand the characteristics of LUs, we considered three different LU metaphors and saw that

- "querying+browsing" requires extensive domain knowledge, similar to that used by librarians in quiding customers,
- "keyword and/or natural language" is not mature yet, as real queries are rich and require natural language, whereas users of existing NL systems do not pose rich queries.
- "immersion (goggles)" is not always feasible because navigation is an important issue; hence, feedback/guidance is again very essential.

We also noted that structured data may be relatively easy to access using this approach.

Consequently, we concluded that probably new metaphors will be necessary to satisfy the retrieval needs of end users. On the other hand, whatever metaphor is used, users habits and capabilities must be taken into account and system must provide guidance to the end user.

The language for developers (LD), on the other hand, does not need to be very easy to use. Consequently, SQL style is probably acceptable. An LD for digital libraries must bring the best properties of information retrieval and databases together. The language must integrate query specification, source description, and user description tasks. An all-encompassing multimedia language standard is probably not realistic; hence, we need to extend existing general purpose languages with standardized multimedia predicates instead of developing standard multimedia/DL languages. Such a language by itself does not need to be complete for all DL applications, but should be extensible through these standard predicates. Most importantly, it has to provide means to support LUs; i.e., it has to describe user characteristics and provide hints and guidance to the end user.

Consequently, the group agreed that an LD should (very loosely) look as follows:

```
SELECT <presentation specifications>
HINT (or GLORIFIED EXPLAIN) <hint + system feedback specifications>
FROM <source descriptions>
FOR <user description>
WHERE <standard predicates + (maybe) some new operators>
```

We also agreed that such a language maybe hard to use; but, since it is not for the end users, but for the developers, we felt that it is probably acceptable.

Specialized Applications of Digital Libraries

Prepared by Elisa Bertino

The goal of this working group was to discuss requirements and functions for DL's in different application areas, including medicine, GIS, humanities, arts, laws, bibliography and reference materials. DL management systems appear still to be very much application-dependent and their development is very expensive. In order to alleviate this problem, it is important to assess whether some common architectural framework can be devised, so that reuse of design and components becames possible.

In order to better organize the discussion, the working group has identified the following relevant functions that a DL system should carry.

- **Data Acquisition.** Any DL needs to acquire data. Data can be acquired from external sources or can produced internally within the DL itself. The former case often requires human intervention and can be quite expensive. Also, for some applications (see as an example the Heraldry DL presented by Wolf-Tilo Balke), application-dependent knowledge may be needed to effectively acquire data. The latter case may require less human intervention to acquire the data, because they are already represented according to models or formats of the DL itself. A relevant issue here, however, is how to improve the process of new data generation. In this respect, re-use (as today intended in object-oriented methodologies) may be very relevant.

- **Knowledge Enrichment.** Acquiring data in general is not enough to enable an effective use of these data by users. Therefore, in most DL's, data are enriched by associating some knowledge to them. Such a knowledge can be introduced in a variety of forms, including indexing, concept associations and classification, correlation and links. This process of knowledge enrichment is usually application dependent, requires domain experts, and also depends on the specific media on which data are recorded. Also, for some applications, this process requires an organizational structure, in charge of approving for example the introduction of new concepts or classification categories.

- **Information Querying.** Such function is the most relevant since it is related to how the users interact with the DL in order to retrieve information they need. In general, DL's for different applications differ quite a lot in this respect. Very often different information querying modes exist even within the same DL. However, most applications seem to require the capability from the DL to take into account user preferences and profiles and also ranking functions.

- **Information Presentation.** Information presentation deals with tools and techniques for delivering the retrieved information to the users. Such function is crucial in order to gain high user acceptance of the system. A DL system should be able to provide different qualities of services and must have provision for supporting

universal accesses, so that data can be presented on a broad range of devices including mobile devices, phones, faxes, etc. Techniques such as summarization and abstracting can be also useful, especially for large size DL data objects. Finally, user preferences and profiles should play a relevant role also in information presentation.

- **Information Protection.** Such function actually encompasses a number of different functions, dealing with securing the rights of the information producers and users as well as to enforce possible security policies of the organization owning the DL. Relevant functions for securing the rights of the information producers include intellectual property protection and copyright. Securing the rights of the users implies supporting anonymity as well as ensuring the privacy of data concerning user preferences and profiles. Organizational security policies require the development of adequate access control mechanisms enabling the specification of which users can access which data objects for which purposes. Finally, a relevant protection aspect is related to the use of secure payment systems, through various forms such as subscription, small billing charges, pre-payment, credit cards. Most DL systems need some form of protection, often encompassing many of the above functions. This is an area where it is probably possible to develop some common mechanisms for use across a variety of different application domain DL's.

5 Other Participants

- Peter M.G. Apers, University of Twente
- Elisa Bertino, University di Milano
- Alejandro Buchmann, TU Darmstadt
- Klaus Dittrich, Universität Ziurich
- Andreas Heuer, Universität Rostock
- Werner Kießling, Universität Augsburg
- Niels Nes, CWI - Netherlands
- Kai Pollermann, TU Braunschweig
- Gunter Saake, Universität Magdeburg
- Kai-Uwe Sattler, Universität Magdeburg
- Eike Schallehn, Universität Magdeburg
- Bernd Walter, Universität Trier
- Yelena Yesha, University of Maryland