

Methodologies for the Long-Term Preservation of and Access to Software-Dependent Electronic Records

PURPOSE & GOALS

The San Diego Supercomputer Center (SDSC), an organized research unit of the University of California, San Diego, has developed expertise with a variety of electronic records and technologies projects and requests a grant from the National Historical Publications and Records Commission (NHPRC):

- to conduct research on the long-term preservation of and access to software-dependent data objects, *and*
- to develop prototypes that will lead to the creation of useful tools for archivists to preserve and provide access to electronic records over the long-term.

Motivation

The area of research we propose deals with the retention of software-dependent electronic records. The purpose of this work is to determine methods for preserving such information and build on the emerging body of work on the preservation of electronic records.

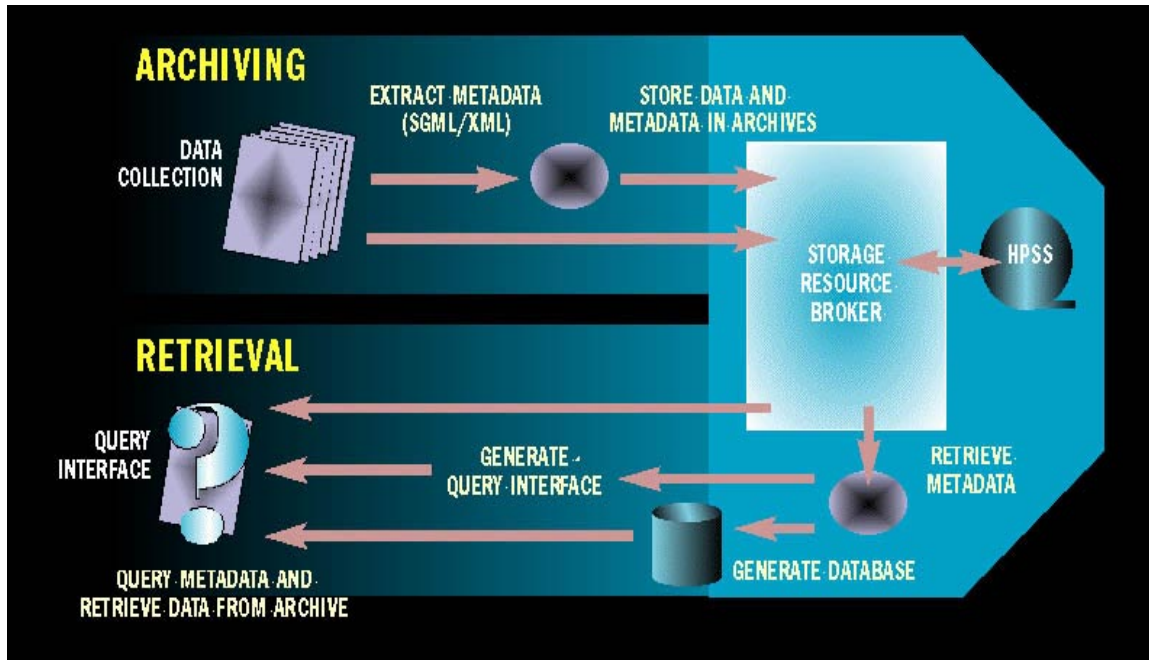
Electronic information such as E-mail messages with attachments, spreadsheet and word processor files, Geographic Information System (GIS) data, database management systems (DBMS) records, and Web pages (e.g., proceedings of the Congress) present unique challenges for archivists.

Background

The San Diego Supercomputer Center is well versed with aspects of system and media evolution and periodically migrates all of its electronic records to new formats and new systems. This activity, while still feasible, ends up consuming vast amounts of resources and staff time. SDSC is developing a framework for persistent archives that will handle a variety of native formats and scale as the information grows. Some of the lessons learned are:

- ❑ **Information input scaling**
 - ❑ Amount of information is growing exponentially
- ❑ **Storage infrastructure scaling issues**
 - ❑ Implies archiving the new data can be the major task
- ❑ **Maintain ability to retrieve information from a collection over 400 years**
 - ❑ The technology used to instantiate the collection changes every 3 years
 - ❑ The technology used for data presentation changes every 4 years
 - ❑ Technology used to archive the collection changes every 5 years

The implication is that there is a significant need for scalable infrastructure and an infrastructure independent description of electronic collections.



The above figure sketches a life cycle architecture for the preservation of digital records that we have begun to develop, based on the analysis of various digital collections from NARA (See Appendix B).

Goals

The organizing principles we adhere to are:

- ❑ Electronic records need to be input as infrastructure independent digital objects
- ❑ Context information about relationships (collection information) needs to be maintained
- ❑ The structure of a collection can be derived through a learning process
- ❑ New electronic records can then be validated against the collection structure
- ❑ Error handling mechanisms can be developed
- ❑ Mechanisms to handle the natural evolution of a collection's structure must be developed

We view the preservation of software-dependent records as a two-fold problem. The first aspect is the preservation of the original record and making it available at any later time. The second is creating a digital representation of the content of the electronic record in a software independent fashion and use this digital object as an infrastructure independent proxy (IIP) for the original record.

We propose to demonstrate that such a digital object can be created based on the analysis of 3 classes of electronic records (text, compound, spatial), and that sustained access to the digital object can be provided through software tools. We further propose to demonstrate a mechanism to partially automate the process of digital object creation and management, and to demonstrate a prototype of such an infrastructure independent management tool in the form of an Archivists' Workbench (AW) software package. We will investigate issues of robustness, structural change and scalability.

One of several approaches we are investigating is the use of markup standards like SGML or XML (<http://www.w3.org/XML/>). The current use of XML indicates the possibility of regarding distributed sources of electronic records as semistructured information. The advantage of modeling electronic records as semistructured information is that such a representation can take advantage of the structure in the record, and simultaneously allow the flexibility to tolerate some variations in individual elements of the record. The second advantage is that it allows us to take advantage of the emerging standard of information representation and exchange on the World Wide Web, the primary vehicle of sustained information access for years to come. This choice further allows us to build on the growing body of research that aims to integrate multiple sources of distributed semistructured information. Included in this research effort is our current work at

UCSD/SDSC on *wrapper-mediator* systems based on XML. A wrapper is a piece of software that acts as a translator between the native format of an information source and a commonly agreed protocol (XML for us). The end-user or application interacts with a piece of software called mediator that collects information from multiple wrappers and allows the user to query the underlying information sources.

In this project we expect to utilize and extend our research results in the following manner:

- We build on the premise that text-encoding like ASCII or Unicode and bitmap-encoding of images are infrastructure independent.
- We contend that tagged representation of structured information in the style of XML is an access-friendly infrastructure independent representation of self-describing records
- Our method of IIP creation is to develop wrappers around software outputs such that
 - All metadata describing the context of the document is converted to an XML document with a well-defined DTD (Document Type Definition) (<http://www.w3.org/XML>)
 - all textual information is converted to XML documents
 - all pictures are converted to bitmaps
 - all references to pictures and to other documents in a record are converted to persistent links, which are also represented in an XML-compliant form
- We provide long-term retention of the IIP along with the original with appropriate referential integrity enforced between them.

SIGNIFICANCE

The proposed project directly addresses one of the top-priority goals in the NHPRC Strategic Plan:

Goal 3: *The NHPRC will enable the nation's archivists, records managers, and documentary editors to overcome the obstacles and take advantage of the opportunities posed by electronic technologies by continuing to provide leadership in funding research-and-development on appraising, preserving, disseminating and providing access to important documentary sources in electronic form.*

Also, our project directly addresses the recommendations in *Research Issues in Electronic Records*. This document identifies four categories of activities that archivists need to undertake

to develop successful methods for the management of archival electronic records: (1) analysis, (2) advocacy, (3) action (later called “*basic program development*” in the 1995 Electronic Records Grant Suggestions” document), and (4) research & development. The R&D component establishes a framework of ten questions, the first three of which are:

1. *What functions and data are required to manage electronic records in accord with archival requirements? Do data requirements and functions vary for different types of automated applications?*
2. *What are the technological, conceptual, and economic implications of capturing and retaining data, descriptive information, and contextual information in electronic form from a variety of applications?*
3. *How can software-dependent data objects be retained for future use?*

Our project primarily focuses on the third question. Research on the retention of software-dependent data objects is paramount in order to build a strong foundation for the further development of requirements and policies for the long-term preservation of and access to electronic records.

Finally, our project will build on the lessons learned from an on-going one-year project between SDSC and NARA called the “*NARA DOCT / Electronic Records Management Project*” (<http://www.sdsc.edu/NARA/>), started on October 1, 1998. Also see the supplemental information in Appendix B. In this project issues related to data ingestion, persistent storage, and metadata archiving are examined against nine unique digital NARA collections. Early technology demonstrations and tools are being constructed.

This project will attempt to extend the earlier project and test the technology demonstrations and tools for scalability, both scaling up towards the federal collections of the NARA collaboration, and scaling down towards the smaller holdings of state and local governments, college and university archives, and other small repositories, stress-testing methodologies and tools in the process. It is important that the tools developed to provide for the long-term preservation and access to electronic records are as usable in a small repository as they are in state or national archives.

Other projects that are likely to provide valuable insights include: InterPARES, the DoD RMA Certification Program, and UPF:

- **the InterPARES Project (<http://www.interpares.org>) is a research initiative which is developing the theoretical and methodological knowledge required for the permanent preservation of authentic records created in electronic systems. Model strategies, policies and standards are expected to follow.**

This project is expected to produce a taxonomy of electronic records, based on findings from the work on "Conceptual Requirements for Preserving Authentic Electronic Records". Should such a taxonomy and record processing principles be developed, we expect this to influence our XML namespace design and IIP extraction.

- **The DoD 5015.2-STD, Design Criteria Standard for Electronic Records Management Software Applications (<http://jite-emh.army.mil/recmgt/>) is being enhanced and a test and certification program for vendor software compliance is being implemented.**

We will investigate the applicability of our long-term preservation of and access to software-dependent electronic records techniques to the records produced by such certified software.

- **the Universal Preservation Format, UPF, (<http://info.wgbh.org/upf/>), while developing a self-describing data storage mechanism for moving image files, may lead to useful insights to the permanent preservation of all digital objects including electronic records.**

We will investigate the compatibility and interoperability of any domain-specific preservation formats resulting from this project.

PLAN OF WORK

The project's overall goal is to research key functions of an Archivists' Workbench (AW) and to prototype them using different classes of software-dependent electronic records.

The categories of software-dependent records will span the following range:

- **a. Textual records:** These records are primarily ASCII. They can be completely unstructured (like a block of text), unstructured but structurable (records of the proceedings of the US Congress) or semi-structured (HTML).
- **b. Compound records:** The documents are typically products of existing software and can have different formats (pdf, Excel, Word, E-mail) and encodings. We call them “compound” because they often embed different types of objects within them. Typical examples include Microsoft Word documents that can have images in them, or emails that can have non-ASCII attachments.
- **c. Spatial records:** A special class of software dependent records is the data produced by geographic information systems. While many different government agencies produce spatial records, the fact that these records contain geometric objects makes them particularly difficult to preserve in an infrastructure-independent fashion.

The key functions we propose to research are:

- I. **Input:** Corresponding to what has been called "ingestion" in our previous work for NARA, this function entails the task of conversion from a record’s native format to a canonical XML equivalent. Records could be coming from a local file, or over the Web, or they could be produced by some other software program such as a GIS or DoD-certified software for electronic records management and organization.

The **research issues** involved here include:

- Where is (or is not) the format conversion process lossless?
- If the conversion is lossy, does it affect the quality of the IIP?
- For embedded objects, how should the component objects be “linked” in the infrastructure-independent representation so that they can be “stitched back” together when a user accesses the record?
- For spatial information, what is the appropriate tagged representation of geometric information?

Some of the **development aspects** of the project will include:

- Conversion of "a." records from HTML to XML

- Conversion of "b." records from native format to HTML to XML
- Conversion of "c." Records such as ArcView Shapefile to SDTS to XML. SDTS stands for Spatial Data Transfer Standard (see <http://mcmcweb.er.usgs.gov/sdts/>)

II. DTD Creation: In our prior work, DTDs have been created manually by an information expert. Based on the current literature in information source wrapping, we believe that for several classes of electronic records, this process can at least be partially automated by inspecting several example records and developing a “model” DTD. Subsequently this model will be refined to reconcile minor variations that can appear in the structure of the electronic records in further examples. As the “learning” process stabilizes, the system will have a fairly robust DTD.

Research issues include:

- What kind of learning algorithms can be applied for each class of records?
- When do these algorithms fail and why?
- How can we achieve a balance between automatic and manual efforts in creating a robust DTD in a reasonable time, using a reasonably small number of examples?

The **development aspects** of this task primarily include:

- experimentation with techniques proposed in the literature and possible advancement of these techniques
- incorporation of promising techniques in the prototype AW

III. Document Validation: This refers to the parsing of a valid XML document according to a related DTD. Fortunately, a number of software products already exist in the market to achieve this. However, their performance needs to be evaluated.

Our **evaluation** will be based upon factors like:

- Do these systems accept XML namespace definitions? This is important because standards such as SDTS can be accommodated into an XML document by creating a suitable namespace.

- How accurate is their validation?
- What kind of error reports do they produce for invalid documents?
- How easily can they be integrated with the rest of the AW prototype?

The **development** process will involve:

- Integration of existing software with AW prototype
- Development of additional error checking if required by the next stage
- Integrating a valid document with the regular electronic records ingestion process currently being developed for NARA

IV. Error Handling: When the validation process encounters errors, a variety of error recovery mechanisms may be initiated. In any case, an error log will be generated, along with some statistics of the type and frequency of errors. The record can be rejected and sent back to the source, either locally or through the Web or by E-mail. Alternately, it may be passed, but annotations showing the errors may be associated with the record. As a variant, some records may be passed conditionally.

Research issues will include:

- What kinds of errors normally occur in practice?
- What errors or inconsistencies cannot be detected by the previous process?
- What additional means are needed to detect these semantic errors?
- How does the archivists set up the rules specifying the error recovery procedure?

The **development process** will include:

- Development of the error handling suite
- Development of the error annotation mechanism in the case an erroneous result is stored
- An interface for the archivist to define the policy of error management

V. DTD Evolution: Long-term preservation must account for possible changes in the underlying collection DTD. While ad-hoc methods can be applied to accommodate for

DTD evolution, this is primarily a research topic. We expect to use the research output of our colleagues in the department of Computer Science at UCSD and those found in literature to suggest and experiment with some plausible approaches toward the problem.

Some of the **research issues** are:

- How do we know that the DTD needs to be updated?
- Since an evolved DTD is first going to appear as an anomalous instance, is it possible to discover the need to evolve from statistics of error occurrence?
- To what degree can we automate the process of inferring the updated DTD?
- How do we manage collections that have multiple (older and newer) DTDs? How do we provide uniform access to record instances regardless of their DTD version?

We present these tasks as a table, organized by the type of record and the function to be accomplished. Note that the tasks get increasingly complex along the rows and the columns of the table.

	a. text records	b. compound records	c. spatial records
I. Input	<i>I.a.</i>	<i>I.b.</i>	<i>I.c.</i>
II. DTD Creation	<i>II.a.</i>	<i>II.b.</i>	<i>II.c.</i>
III. Document Validation	<i>III.a.</i>	<i>III.b.</i>	<i>III.c.</i>
IV. Error Handling	<i>IV.a.</i>	<i>IV.b.</i>	<i>IV.c.</i>
V. DTD Evolution	<i>V.a.</i>	<i>V.b.</i>	<i>V.c.</i>

Advisory Group: While the proposed project will build on our previous and current work with NARA, we also expect to develop an archival advisory group. This group will consist of expert archivists providing feedback on:

- The metadata attributes necessary for long-term preservation of different classes of records
- The nature of relationships to be maintained across records and collections
- The usability of the tools developed in the course of the project

We will communicate with the Advisory Group via electronic means (E-mail, Web, other).

Comment: Although we will not request additional funds, we expect to include, in the Archivists' Workbench, relevant tools that we develop for other Government funded projects, such as the National Partnership for Advanced Computational Infrastructure (NPACI) from the National Science Foundation (NSF) (see <http://www.npaci.edu>).

PRODUCTS, PUBLICATIONS, and OUTCOMES

By pursuing activities for

- conducting research on the long-term preservation of and access to software-dependent data objects, *and*
- developing prototypes that will lead to the creation of useful tools for archivists.

this project will help validate and develop methodologies and tools that have the potential to be useful to the archival community.

By addressing spatial data (GIS information), this project will contribute to the further development of the National Spatial Data Infrastructure.(NSDI) by looking at the applicability of the Spatial Data Transfer Standard (SDTS) and the Content Standard for Geospatial Metadata (CSDGM) for the Archivists' Workbench (AW).

The primary products of this project will be prototype software that will be made publicly available, reports with the findings and recommendations of our research, and the publication of results in suitable archival and technical journals.

The software tools for the Archivist's Workbench that are ultimately made publicly available will depend in large part on the findings of our research, but we anticipate that the software will include the following tools:

- Format conversion software for text, HTML, compound documents, and GIS records
- Tools for example-based DTD learning
- Document validation and error reporting module
- Error handling modules
- Specification tool for error management policies
- DTD evolution tools

We will prepare reports of the findings and recommendations of the project and make them widely available through the Internet. In addition we will send notices of the availability of reports and requests for comments to the appropriate listservs.

We intend to submit articles to the *American Archivist*, *Communications of the ACM*, and *ACM Transactions on Information Systems* for publication. We will acknowledge NHPRC support in all products and presentations.

PERSONNEL

The research and development performed at UCSD/SDSC is a team effort and leverages benefits of the prior work performed by scientist and programmer colleagues at this institution. The tasks to be performed uniquely for the proposed project will, however be conducted by the following personnel.

Project coDirector: Amarnath Gupta (20%) will be responsible for the following aspects of the project:

- (a) analysis of the software formats under investigation
- (b) design of format conversion and DTD extraction process
- (c) design of the DTD learning algorithm
- (d) overall management of the project including the software demonstrations planned

Project coDirector: Richard Marciano (20%) will be responsible for the following aspects of the project:

- (a) extraction of DTD for spatial information
- (b) design of the Archivists' Workbench software architecture
- (c) error handling in validation
- (d) management of the storage process

The scientists will be jointly responsible for the DTD evolution research.

A graduate research assistant (TBA 50%) will be hired from the department of Computer Science, UCSD, through the usual student recruitment process of the university. The student will be required to have a background in database systems, Java programming, and XML manipulation. The student will be primarily responsible for implementation of the software Archivists' Workbench. It is expected that part of the student's responsibility will be to investigate the theoretical and operational feasibility of alternate approaches of achieving the same task.

The successful development of this project will rely on the expertise of two principal researchers at UCSD and SDSC. Both Amarnath Gupta and Richard are members of the DICE and MIX research groups (<http://www.npaci.edu/DICE>).

DICE, *Data-intensive Computing Environments*, is an effort that is looking at building a national digital library that accelerates the publication of scientific data. This requires the integration of distributed persistent digital archives, hierarchical storage systems, databases, data-handling systems, and digital libraries into scientific information repositories.

MIX, *Mediation of Information Using XML*, is a collaboration between several DICE researchers and the UCSD Database Lab in which we are developing wrappers for a variety of information sources including, relational databases, GIS systems, and Web sites with HTML pages. Mediation is based on the *MIXm* mediator--and the associated XMAS query language--being developed by the UCSD Database Lab component of MIX project.

Amarnath Gupta, is particularly interested in mediating spatiotemporal information. Spatiotemporal information stored in GIS, spatial databases and image databases is central to many application domains such as environmental monitoring, regional planning and crime mapping. However, for any large-scale application, what is needed is the ability to combine information from several independent spatiotemporal information sources that may be distributed, running on heterogeneous platforms, and developed with a variety of software tools. Gupta and Marciano are developing tools to enable integration of diverse spatiotemporal sources through an XML-based information interchange protocol. These tools include a "wrapper" designed to export information from spatiotemporal information sources, so that they appear as XML document sources to an application, and a "mediator", which performs distributed query processing from "wrapped" spatiotemporal information sources.

Richard Marciano is interested in identifying the information architecture required to build and maintain a viable persistent archive. The approach is based upon the concept that both the original digital objects and the information required to assemble the digital objects into a data collection must be preserved. Digital objects are not stored as stand-alone entities, but instead are preserved as members of a digital data collection. Persistence is achieved through identification of metadata for all attributes related to digital object properties and collection organization. Persistence is demonstrated by dynamically building the data collection from the individual data objects stored in the archive, dynamically creating the queries needed to discover information within the data collection, and dynamically constructing the presentation interface for the digital objects discovered through a query. The approach that is being followed is broken down into four areas corresponding to the automated loading of data and generation of metadata, the long-

term storage of digital objects, support for information discovery against the archived data, and the presentation of the discovered data objects.

SUGGESTED REVIEWERS

Larry Brandt

CISE/EIA
National Science Foundation
4201 Wilson Boulevard
Arlington, Virginia 22230
(703) 306-1981 (phone)
(703) 306-0589 (fax)
lbrandt@nsf.gov

Peter Bloniarz

SUNY Albany
Center for Technology in Government
1535 Western Ave.
Albany, NY 12203
(518) 442-3892 (phone)
pbloniarz@ctg.albany.edu

Terry Smith

Project Director, Alexandria Digital Library
Departments of Computer Science and
Geography
University of California, Santa Barbara
Santa Barbara, CA 93106
(805) 893-2966
smithtr@cs.ucsb.edu

Hector Garcia-Molina

Department of Computer Science
Stanford University
Stanford, CA 94305-9040 USA
Phone: (650) 723-0685
Fax: (650) 725-2588
hector@cs.stanford.edu

Andreas Paepcke

Stanford University
Gates Computer Science, RM 426
Stanford, CA 94305
(650) 723-9684 (phone)
(650) 725-2588 (fax)
paepcke@cs.stanford.edu

Merritt Jones

Vice-Chair, MSSTC
The MITRE Corp.
Mail Stop: Z-250
7525 Colshire Drive
McLean, VA 22102-3481
(703) 883-5471 (work)
(703) 883-3615 (fax)
merritt@mitre.org

Richard Watson

Lawrence Livermore National Laboratory
P.O. Box 808, L-66
7000 East Avenue
Livermore, CA 94550
(510) 422-9216 (work)
(510) 423-4820 (fax)
dwatson@llnl.gov

APPENDIX A: CVs

Amarnath GUPTA

University of California, San Diego
San Diego Supercomputer Center
9500 Gilman Drive
La Jolla, CA 92093-0608
619-822-0505
619-534-8380 (fax)
amarnath@sdsc.edu

EDUCATION

- 1994 Ph.D. (Engineering) in Computer Science, Jadavpur University, India
1987 M.S. in Biomedical Engineering, University of Texas, Arlington
1984 B.Tech. in Mechanical Engineering, Indian Institute of Technology, Kharagpur, India

PROFESSIONAL EXPERIENCE

- 1998-present Asistant Research Scientist, San Diego Supercomputer Center
1995-1998 Scientist, Virage, Inc., San Mateo, CA
1994-1995 Sr. IT Engineering, CMC Limited, India
1988-1994 Computer Engineer, Indian Statistical Institute, Calcutta, India
1984-1987 Research Assistant, Biomedical Engineering, University of Texas, Arlington

RELATED PUBLICATIONS

C. Baru, A. Gupta, B. Ludäscher, R. Marciano, Y. Papakonstantinou, P. Velikhov, A. Yannakopoulos, "XML-Based Information Mediation with MIX". Proceedings of the SIGMOD'99 (submitted).

A. Gupta and C. Baru, "An Extensible Information Model for Shared Scientific Data Collections", Future Generation Computer Systems, Summer 1999 (to appear).

A. Gupta, R. Marciano, I. Zaslavsky, C. Baru, "Integrating GIS and Imagery through XML-Based Information Mediation", NSF International Workshop on Integrated Spatial Databases: Digital Images and GIS, June 1999.

Gupta A., Santini, S., Jain, R., In search of information in visual media, Communications of the ACM, vol. 40, 12, pp. 35-42, December 1997.

Katkere, A., Schlenzig, J., Gupta, A., Jain, R., Interactive video on the WWW: Beyond VCR-like interfaces, Computer Networks and ISDN Systems, vol. 28, no. 7-11, pp. 1,559-1,572, May 1996.

OTHER RELEVANT PUBLICATIONS

Gupta, A., Hampapur, H., Gorkani, M., Jain, R., On summarization of video, International Conference on Image Processing, Santa Barbara, CA, 1997.

Gupta, A., Weymouth, T., Jain, R., Semantic queries in image databases, IFIP Transactions A (Computer Science and Technology), vol. A-7, pp. 201-215, 1992.

Gupta, A., Jain, R., Visual information retrieval, Communications of the ACM, vol. 40, 5, pp. 70-79, May 1997.

Defining objects in multi-camera video databases, Proc. Storage and Retrieval for Still Image and Video Databases IV, San Jose, CA, vol. 2670, pp. 120-131, February 1-2, 1996.

Gupta, A., S. Moezzi, A. Taylor, S. Chatterjee, R. Jain, M. Goldbaum, S. Burgess, Content-based retrieval of ophthalmological images, Proc. International Conference on Image Processing, Lusanne, Switzerland, vol. 3, pp. 701-704, 1996.

COLLABORATORS

Ramesh Jain, Department of Electrical and Computer Engineering, UC San Diego

Gio Wiederhold, Department of Computer Science, Stanford University

GRADUATE STUDENTS

Deboprotim Ghosh, Indian Statistical Institute

Krishnendu Mukherjee, Indian Statistical Institute

Nirupam Sarkar, Indian Statistical Institute

POSTDOCTORAL RESEARCHERS

None

GRADUATE ADVISORS

Aditya Bagchi, Indian Statistical Institute

Anup K. Bandyopadhyay, Jadavpur University

Richard MARCIANO

University of California, San Diego
San Diego Supercomputer Center
9500 Gilman Drive
La Jolla, CA 92093-0608
(619) 534-8345
(619) 822-0906 (fax)
marciano@sdsc.edu

EDUCATION

- 1992 Ph.D. Computer Science, University of Iowa, specialization in Parallelizing Compilers
1989 M.S. Computer Science, University of Iowa
1986 B.S. Electrical Engineering & Avionics, National School of Civil Aviation, Toulouse

PROFESSIONAL EXPERIENCE

- 1996-present Research Scientist, San Diego Supercomputer Center (SDSC)
1995-1996 Associate Staff Scientist (Environmental Science), SDSC
1994-1995 Computational Scientist, National Supercomputing Center for Energy & the Environment UNLV, Las Vegas, Nevada
1993-1994 Research & Development Consultant, Supercomputing & Parallel Processing Lab, Geography Department, University of Iowa
1986-1992 IBM Fellow, Supercomputing Consultant, Research Assistant, Parallel Programming Instructor, University of Iowa.

RELATED PUBLICATIONS/PRESENTATIONS

“XML-Based Information Mediation with MIX”, C. Baru, A. Gupta, B. Ludäscher, R. Marciano, Y. Papakonstantinou, P. Velikhov, A. Yannakopoulos, Proceedings of the SIGMOD'99 (submitted).

"Integrating GIS and Imagery through XML-Based Information Mediation", A. Gupta, R. Marciano, I. Zaslavsky, C. Baru, NSF International Workshop on Integrated Spatial Databases: Digital Images and GIS, June 1999.

"Integrating Information with XML", C. Baru, R. Marciano, Presented at the April Internet2 meeting, April 27-29, 1999, Washington, DC.

"Metadata to Support Information-Based Computing Environments", C. Baru, R. Frost, R. Marciano, R. Moore, A. Rajasekar, M. Wan, *2nd IEEE Conference on Metadata (METADATA '97)*, September, 1997, Greenbelt, MD.

"Towards the Interoperability of Web, Database, and Mass Storage Technologies for Petabyte Archives", R. Moore, R. Marciano, M. Wan., T. Sherwin, R. Frost, *Procs. Fifth NASA Goddard Space Flight Center Conference on Mass Storage Systems and Technologies (MSS'96)*, March 1996.

OTHER RELEVANT PUBLICATIONS

"Discrete-Event Simulation Studies of Archival Storage", W. Schroeder, R. Marciano, SDSC White Paper, April 1999.

"Analysis of HPSS Performance Based on Per-File Transfer Logs", W. Schroeder, R. Marciano, J. Lopez, M. Gleicher, G. Kremenek, C. Baru, R. Moore, *Procs. Seventh NASA Goddard Conference on Mass Storage Systems & Technologies, and Sixteenth IEEE Mass Storage Systems Symposium, March 15-18, 1999, San Diego, CA.*

"Rear-Projecting Virtual Data onto Physical Terrain", D. Clark, R. McKeon, R. Marciano, M. Bailey, *IEEE Visualization '98, Oct. 18-23, 1998, Research Triangle Park, N.C.*

"Temporal & Spatial Data Prototype: 100 Years of Neighborhood Urban Change in Mission Hills", R. Marciano, R. McKeon, L. Cruse, 1998 Annual ESRI GIS Conference. A poster generated from this work received the "Most Artistic" award.

"Massively Parallel Strategies for Local Spatial Interpolation", M. Armstrong, R. Marciano, *Spatial Data Handling, International Symposium on Spatial Data Handling, Vancouver, Canada, July 12-15, 1998.*

APPENDIX B:

MIX: Mediation of Information Using XML

(1 sheet)

NARA/SDSC: Electronic Records Management & Persistent Archives

(1 sheet)