

PROJECT SUMMARY

1. PURPOSE AND GOALS OF PROJECT

The project will:

- conduct research on the long-term preservation of and access to software-dependent data objects, *and*
- 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.

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 IIP objects 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.

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.

2. SIGNIFICANCE AND RELATIONSHIP TO NHPRC GOALS AND OBJECTIVES

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* and focuses on question (3). 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.

3. PLAN OF WORK FOR GRANT PERIOD

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: textual, compound, and spatial. The key functions we propose to research are: input or ingestion, DTD creation, document validation, error handling, DTD evolution.

Milestones:

- Year 1. Development of technology. Initial demonstration of technology using files with textual materials
- Year 2. Demonstration of initial version of archivists' workbench software, with DTD learning and validation. Demonstration of long-term preservation of pdf and files with mixed textual/graphical content.
- Year 3. Final Demonstration of software. Demonstration of GIS record preservation using collections from representative archival repositories. We have had initial discussions with researchers from university and state archives. We also plan to use publicly available information from the Federal Geographic Data Committee (FGDC) and the San Diego Association of Governments (SanDAG) (see <http://fgdc.er.usgs.gov/> and <http://www.sandag.cog.ca.us>).

We also expect to develop an archival advisory group consisting 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 and the usability of the tools developed in the course of the project

4. PRODUCTS/PUBLICATIONS TO BE COMPLETED DURING GRANT PERIOD

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 tools for 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; and 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.

5. KEY PERSONNEL

Project coDirector: Amarnath Gupta (20%)
Telephone: (619) 822-0994
Address: University of California at San Diego
San Diego Supercomputer Center
9500 Gilman Drive
La Jolla, CA 92093-0505
Email: gupta@sdsc.edu

Project coDirector: Richard Marciano (20%)
Telephone: (619) 534-8345
Address: niversity of California at San Diego
San Diego Supercomputer Center
9500 Gilman Drive
La Jolla, CA 92093-0505
Email: marciano@sdsc.edu

Graduate Research Assistant (TBA 50%)