# Final Report January 2004

## High Performance Interoperable Digital Libraries in the Open Archives Initiative

to the Consejo Nacional de Ciencia y Tecnología (CONACYT) and National Science Foundation (NSF)

Prof. Dr. Edward A. Fox,
Virginia Polytechnic Institute and State University (Virginia Tech), Dept. of Computer Science, Blacksburg, Virginia, USA

Prof. Dr. J. Alfredo Sanchez
Universidad de las Américas, Puebla, México

Prof. Dr. David Garza-Salazar,
Instituto Tecnológico y de Estudios Superiores de Monterrey, México

# 1. Participants

## 1.1. What people have worked on the project?

### Virginia Polytechnic Institute and State University, Blacksburg, USA

- Prof. Dr. Edward A. Fox
- W. Ryan Richardson
- Ashwini Pande
- Malini Kothapalli
- Nigel Carty
- Malar Chinnusamy

### Universidad de las Américas, Puebla, México

- Prof. Dr. J. Alfredo Sanchez
- Nohema Castellanos
- Lourdes Fernandez
- Yazmin Morales
- Sandra E. Nava
- Christian Rebollar
- Natalia Reyes

**Instituto Tecnológico y de Estudios Superiores de Monterrey, México**

- Prof. Dr. David Garza-Salazar
- Adán Salinas
- Carlos Medina
- Carlos Rivero
- Juan Córdova Briones

# 1.2. What other organizations have been involved as partners?

- None.

# 1.3. Have you had other collaborators or contacts?

We worked closely with the Internet2 Distributed Storage Infrastructure (I2-DSI) project and the Internet Backbone (IBP) project, based at the University of Tennessee-Knoxville.
We collaborated with other researchers at Virginia Tech including but not limited to Hussein Suleman, Jun Wang, Marcos André Gonçalves, and Aaron Krowne. Also we collaborated with two of Dr. Fox's class groups, one consisting of Ashwini Pande and Malini Kothapalli, and the other of Nigel Carty and Malar Chinnusamy.
There is also a general collaboration with various participants in the international Open Archives Initiative.
We made contact with Ramon Suarez Espinosa, Head Librarian at the Biblioteca Central at the Universidad Autónoma de Chapingo.

# 2. Activities and Findings

# 2.1. What were your major research and education activities?

The scope of this project is high performance mechanisms for interoperable distributed digital repositories. We apply Open Archives Initiative ideas and concepts to the storage and retrieval of electronic theses and dissertations (ETDs), and work to make these more available to students by means of visualization tools. Towards these ends we are building OAI-based digital library services and software tools:

- MARIAN digital library system

- U-DL-A digital library system
- Phronesis digital library system
- UVA visualization interface
- Envision visualization interface
- Sophia personal classification system for OAI archives
- VIDI protocol to connect visualizations to digital libraries

We also developed a prototype system that allows us to replicate an OAI archive over the Internet2 Distributed Storage Infrastructure (I2-DSI), using what is called a channel in I2-DSI terminology. This provides us with both automated mirroring and network layer redirection to help in our goal of providing high performance digital library services. The prototype was used to mirror the contents of Virginia Tech's ETD collection.

UDLA has been continuing their work on U-DL-A, which is an agent-based digital library system. They installed a copy of their MAIDL system on a Sun server at Virginia Tech, making VT one of multiple sites where U-DL-A agents can travel to find documents. They worked on extensions to the OAI Protocol for Metadata Harvesting (OAI-PMH) to support full text searching and other functionality not provided by the base protocol. In addition, UDLA made their online electronic thesis collection http://mailweb.udlap.mx/~tesis/ an OAI archive, and converted it to version 2.0 of the OAI-PMH with assistance from Hussein Suleman.

ITESM continued development of their Phronesis DL system. Its current version, v1.3, supports federated search via the Z39.50 protocol. They added support for federated search via OAI in addition to Z39.50, and they also ported the system from using CGI to a full Java implementation.

Jun Wang at Virginia Tech developed a protocol to allow visualization tools (such as Envision and UVA) to connect to different digital library systems. This protocol is based on the format of OAI-PMH, with extensions specific to visualizations. This work was central to Jun Wang's master's degree research. Jun graduated in May 2002.

## 2.2. What are your major findings from these activities?

- The OAI-PMH is very flexible and can be extended to support full-text searching, visualization, cross-language searching, and digital library implementation.
- MARIAN is a complex DL system whose characteristics (e.g., semantic network structures, weighting schemes, object-oriented DL API) can be extremely powerful in building OAI-based services.

- The monolithic architecture of MARIAN poses a challenge in making its functionality available for other projects. Componentized systems are needed, and these can be achieved using OAI-based protocols such as ODL (Open Digital Libraries).
- Often the biggest impedance to distributed high performance networking solutions is not technology but rather the getting permission to access/reconfigure servers at the various sites. Also marketing plays a role. Some protocols are only supported by certain networking hardware, so if a given site uses another type of hardware, they will not be able to run that protocol. Hence developing standardized protocols is a necessity.
- Through the OAI protocol, it is fairly easy to interconnect very different sources and present them through one easy to use search interface.

## 2.3. What opportunities for training and development has the project helped provide?

We have developed a set of services that will help to increase the availability of student research for scholars. Our services also should improve productivity by students and other researchers by use of visualizations. Multiple courses at Virginia Tech (especially CS5604, Information Storage and Retrieval and CS6604, Digital Libraries) have had one or more project groups learning through involvement in this effort.

## 2.4. What outreach activities have you undertaken?

Multiple tutorials have been given at digital library conferences by Edward Fox, Ryan Richardson, Sandra Nava, and Hussein Suleman. To inform Latin Americans about the Open Archives Initiative, Ryan Richardson gave a presentation at the Amigos Conference, February 21, 2002, in Puebla, Mexico. That same day, Edward Fox gave an invited talk "Digital Library as Infrastructure for Knowledge Management in Academic and Research Institutions: Open Archives Initiative and Educational Demonstrations". The next day he moderated a panel on "Construyendo una plataforma de tecnología integrada para la administración del conocimiento".

A demonstration about MARIAN has been given by Edward Fox and Marcos A. Goncalvez at ECDL 2001 in Darmstadt.

Prof. Edward Fox has organized a workshop about Open Archives at ACM SIGIR'01 in New Orleans.

Alfredo Sánchez moderated a "Mesa de Bibliotecas Digitales" in the CUDI Spring Meeting, Ensenada, Mexico, April 2-4, 2003.

Alfredo Sánchez moderated another "Mesa de Bibliotecas Digitales" in the CUDI Fall Meeting, Puebla, Mexico, October 2003.

A presentation entitled "Programa de bibliotecas digitales de la UDLA" was presented during a day-long videoconference (Día Virtual CUDI), March 14, 2003.

For cooperative work, Edward A. Fox and Ryan Richardson have visited Universidad de las Américas in Puebla in February 2002. Fox gave a tutorial in Spanish on OAI in November 2001 in Montevideo, Uruguay to people who will be trainers about ETD efforts. Fox gave an invited talk Improving Graduate Education: Networked Digital Library of Theses and Dissertations (NDLTD) Round Table on "Digital Theses: A New Vein of Information" at the International Conference on University Libraries: Electronic publishing and library services, on October 23, 2003. The next day, at UDLA, he gave a seminar "Research Overview Related to NSDL". He also spoke in a day-long videoconference (Día Virtual de Bibliotecas Digitales) arranged by UDLA and broadcast around Mexico on Nov. 25, 2003: "Networked Digital Library of Theses and Dissertations (NDLTD)" (in Spanish: La Red de Bibliotecas Digitales de Tesis y Disertaciones).

Ryan Richardson presented "Mirroring an OAI archive with an I2-DSI channel", via videoconference at the Internet2-Distributed Storage Initiative (I2-DSI) conference May 2002.

# 3. Products

## 3.1. What have you published as a result of this work?

### 3.1.1. Major journal publications

Lourdes Fernández, , J. Alfredo Sánchez, A. García, Tales: Integración de tesis en una biblioteca digital avanzada. Scire: Representación y Organización del Conocimiento 8, 2 (Jul.-Dec.), 61-70. Zaragoza, Spain. (ISSN 1135-3761). 2002.

Marcos Gonçalves and Edward Fox. Technology and Research in a Global Networked University Digital Library (NUDL). Ciência da Informação (leading journal of library and information science in Brazil), 30(3): 13-23, Sep./Dec. 2001.

J. Alfredo Sánchez, Sandra Nava Muñoz, Lourdes Fernández Ramírez, G. Chevalier Dueñas. Distributed information retrieval from web-accessible digital libraries using mobile agents. *Upgrade,* Special Issue on Information Retrieval and the Web, 3, 2 (April), 37-43. 2002.

J. Alfredo Sánchez, Sandra Nava Muñoz, Lourdes Fernández Ramírez, G. Chevalier Dueñas. Recuperación de información distribuida de bibliotecas digitales via web utilizando agentes móviles. Novatica 156 (March-April). 21-26. 2002.

J. Alfredo Sánchez, Colecciones digitales universitarias en México. *Biblioteca Universitaria 5*, 2 (July-Dec.), 130-143. 2002.**3.1.1.1**

# Conference/Workshop Proceedings

Nohema Castellanos, J. Alfredo Sánchez, 2003. *PoPS: Mobile access to digital library resources*. Proceedings of the Joint Conference on Digital Libraries (JCDL 2003, Houston, Tex., May).

G. Chevalier Dueñas, J. Alfredo Sánchez, Lourdes Fernández, Sandra Nava, 2001. Viajerus: A framework based on mobile agents for distributed information retrieval. *Proceedings of the Mexican International Conference on Computer Science (Encuentro Internacional de Computación - ENC 01*, Aguascalientes, México, Sept.). 673-682.

Lourdes Fernández, J. Alfredo Sánchez. Community Tales: An infrastructure for the collaborative construction of digital theses repositories. Proceedings of the Sixth International Conference on Electronic Theses and Dissertations (ETD 2003, Berlin, Germany, May).

Marcos André Gonçalves, Paul Mather, Jun Wang, Ye Zhou, Ming Luo, Ryan Richardson, Rao Shen, Liang Xu, Edward A. Fox: Java MARIAN: From an OPAC to a Modern Digital Library System. SPIRE 2002: 194-209

M. A. Medina, J. Alfredo Sánchez, 2002. Agents at the reference desk: Serving information needs and constructing knowledge for wide communities of users. Memorias del XII Congreso Internacional de Electrónica, Comunicaciones y Computadoras (Conielecomp 2002, Acapulco, México, Feb.), 74-78.

Ashwini Pande, Malini Kothapalli, Ryan Richardson, Edward. A. Fox. Mirroring an OAI archive on the I2-DSI channel, Short paper in *Proceedings of 2nd Annual Joint Conference on Digital Libraries*, Portland, Oregon, July 14-18 2002, 293-294.

Carlos Proal, J. Alfredo Sánchez. 2003. Modelado de acervos de bibliotecas digitales. ENC 2003: Avances en Ciencias de la Computación (Sept. 8-12, Tlaxcala, Mexico). 49-54. (ISBN 970-36-0069-7)

N. Reyes-Farfán, J. Alfredo Sánchez. 2003. Personal spaces in the context of OAI. *Proceedings of the Joint Conference on Digital Libraries* (JCDL 2003, Houston, Tex., May).

J. Alfredo Sánchez, Carlos Proal, D. Pérez, A. Carballo. 2001. Personal and group spaces: Integrating resources for users of digital libraries. *Proceedings of the 4th Workshop on Human factors in Computer Systems* (IHC 2001, Florianópolis, Brazil, Oct. 15-17). 183-194.

N. Silva, J. Alfredo Sánchez, Carlos Proal, Christian Rebollar. 2003. Visual exploration of large digital libraries collections. Proceedings of the Latin American Conference on Human-Computer Interaction (CLIHC 2003, Rio de Janeiro, Brazil, August). 147-157.

Hussein Suleman, Ryan Richardson, *Construyendo Bibliotecas Digitales Interoperables: Una Guía Práctica para crear archivos abiertos.* Tutorial presented at the 2nd Biannual Amigos Conference: Cooperación para la administracion del conocimiento (Amigos2002), Puebla, Mexico, February 2002.

Jun Wang, Abhishek Agrawal, Anil Bazaz, Supriya Angle, Edward A. Fox, and Chris North. Enhancing the ENVISION Interface for Digital Libraries. Short paper in Proc. JCDL'2002 Second Joint ACM / IEEE-CS Joint Conference on Digital Libraries, July 14-18, 2002, Portland, pp. 275-276.

## 3.1.2. Books and other one-time publications

Rao Shen, Jun Wang, and Edward A. Fox. A Lightweight Protocol between Digital Libraries and Visualization Systems. In Visual Interfaces to Digital Libraries, eds. Katy Borner & Chaomei Chen, Springer Verlag, LNCS Series, Vol 2539, 2002.

David Garza, Juan Carlos Lavariega, Martha Sordia. Knowledge-based information retrieval and filtering from the web. In Information Retrieval and Administration of Distributed Documents in Internet. Kluwer Academic Publishers. 2003. pp. 53-74.

## 3.2. What web site(s) or other Internet site(s) reflect the project?

The VT digital library page can be found at http:///www.dlib.vt.edu. The MARIAN OAI data provider code is available at http://parsifal.dlib.vt.edu/lib/Marian/. The UDLA project page can be found at http://biblio.pue.udlap.mx/u-dl-a/. The ITESM project page can be found at http://copernico.mty.itesm.mx/~tempo/Projects. More information about MARIAN can be found directly at http://www.dlib.vt.edu/projects/MarianJava/index.html . The OAI page at VT is http://www.dlib.vt.edu/projects/OAI/index.html

General information about the Open Archives Initiative can be found at http://www.openarchives.org/, and dozens of archives can be browsed at http://oai.dlib.vt.edu/cgi-bin/Explorer/oai2.0/testoai.

## 3.3. What other specific products have you developed?

We have built several software tools and packages for development of OAI-based services, including:

### 3.3.1 MARIAN DL system

MARIAN is a multi-user information system designed primarily as digital library infrastructure. It is designed to support large numbers of simultaneous sessions of the sort commonly encountered in library environments: short sequences of often unrelated queries punctuated by browsing and examination of documents. MARIAN also supports query editing and refinement based on an explicit query history. Over the course of the project MARIAN has been converted from C and C++ to Java (in part with National Library of Medicine support) to enhance portability and to support modernization and redesign. It also supports Spanish collections, although we have not much utilized this feature.

### 3.3.2 U-DL-A DL system

U-DL-A is an agent-based digital library system with three objectives:

1. to produce advances in the two major areas in the development of digital libraries, which are constructing digital repositories and enabling services to take advantage of vast digital repositories;
2. to build a digital library that will support graduate and undergraduate education, including: repositories of interest for

students, faculty and researchers, personalized services for accessing (exploring and searching) repositories, environments for communication and collaboration among users, and mechanisms for exchange of information and services with other digital libraries; (This might be called a type of institutional repository, a term recently popularized in connection with D-Space.) and

3. to develop an environment (both technologically and socially) and a testbed to enable research of diverse open issues in the area of digital libraries.

### 3.3.2 Phronesis DL system

Phronesis is a DL system developed at ITESM. It is a multilingual (Spanish/English) DL system based on the MG (Managing Gigabytes) system. It exemplifies the multiple database/gateway approach to DL design, by means of the Z39.50 and OAI-PMH protocols.

### 3.3.3 UVA visualization interface

UVA is based on 3D representations of hierarchical structures to visualize overlapping classification schemes. It allows users to start browsing the library from a default taxonomic point of view, which is represented graphically as a three-dimensional tree. As nodes (representing groups of library items) are selected, taxonomic sub-levels and their relationships with other existing taxonomies are displayed. The user can zoom in and out in this 3D representation, as well as rotate each taxonomic tree, thus keeping a sense of the context in which navigation is taking place. From any node in the 3D trees, users may obtain associated information, such as full bibliographic citations, abstracts, tables of contents, or full documents in a variety of formats and media.

### 3.3.5 Envision visualization interface

Envision is a visualization interface originally developed for the previous version of MARIAN in the mid-1990's. This user-controlled system facilitates examining very large data sets, displaying multiple aspects of the data simultaneously and efficiently, as well as interactive discovery of patterns in the data. The Envision interface was converted to a Java applet as part of Jun Wang's Master's degree work (see Fig. 1 below). However, it was determined that Envision was too closely coupled with the MARIAN system to be of much use in displaying data from other digital libraries. Therefore, many aspects of the Envision system were incorporated in a new Java codebase by Rao Shen at Virginia Tech. Further information on this project is available here http://csgrad.cs.vt.edu/~nkampany/citidel.html.

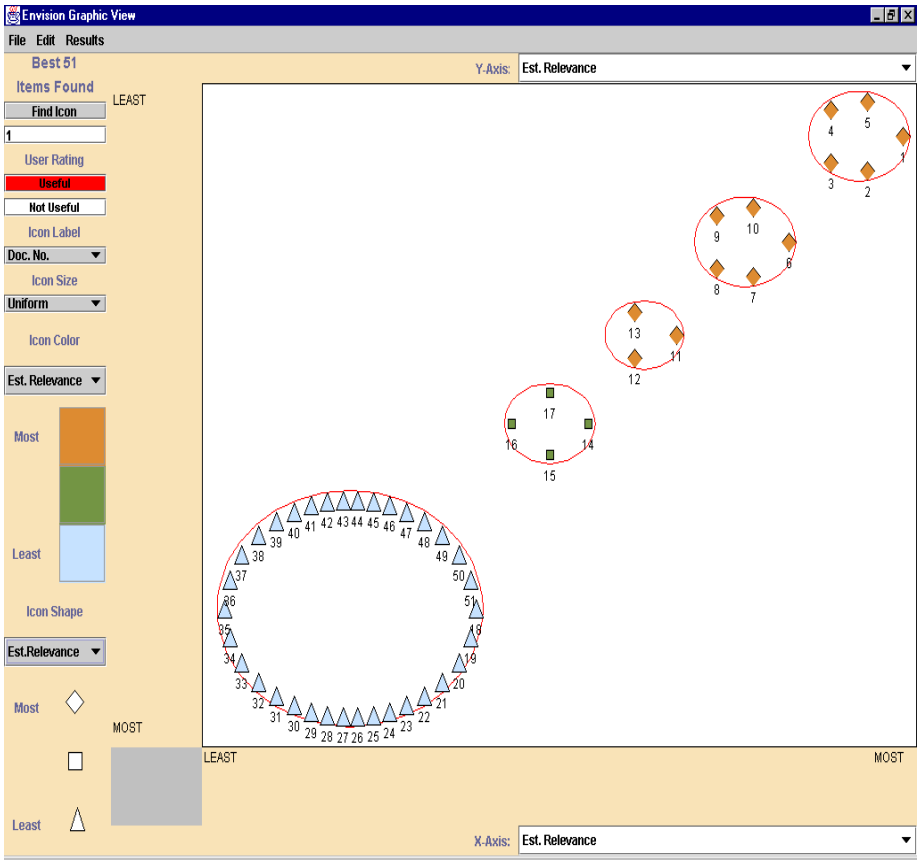


**Figure 1. The revised Envision interface.**

### 3.3.6 Sophia

Sophia is a personal classification system for OAI archives. It is used to facilitate the creation and management of personal classifications of documents in OAI-accessible repositories. It is available at ict3.pue.udlap.mx/rebollar/sophia.html. It was tested at Virginia Tech, demonstrated in several settings, and we gave feedback on it to researchers at UDLA.

### 3.3.7 VIDI protocol to connect visualizations to digital libraries

VIDI is a protocol that was developed to connect visualization tools to disparate digital libraries systems. Essentially it is an extension of the OAI protocol. Jun Wang developed this protocol as part of her Masters degree research at Virginia Tech. The key idea is a precursor to methods that can be implemented in future DL-Viz combinations systems that will be implemented using Web Services.

# 4. Contributions

## 4.1. To the development of the principal discipline(s) of the project?

The mission of the Open Archives Initiative is to promote interoperability, efficiency, flexibility, and scalability of digital library services through the use of a simple, lightweight protocol. We have demonstrated, in a small scale, the applicability of such concepts to build high quality services employing mirroring and visualization.

## 4.2. To other disciplines of science and engineering?

By design, the efforts on this project should serve as a model to apply similar techniques/methodologies to build interoperable information services in other science and engineering areas as well as other organizational levels (by country, by topic, etc.). We have applied our methods to: general theses and dissertations, Computer Science, and medical information (in conjunction with NLM/ORISE support.)

## 4.3. To the development of human resources?

We have introduced OAI to librarians, computer scientists, and students in Latin America, where previously it was largely unknown. We have involved students in multiple classes at Virginia Tech, who now have knowledge of these concepts and technologies. We have involved more than 25 people in the Digital Library Research Laboratory at Virginia Tech in discussions of project activities. We have helped train many people around the world through tutorials, presentations, and visits.

As already mentioned in Section 2.4, two training workshops have been held in Mexico especially for librarians in order to inform about the Open Archives Initiative.

## 4.4. To physical, institutional, and information resources that form the infrastructure for research and education?

We have presented a prototype information service (called a “channel” in I2-DSI terminology) to provide high-speed, redundant access to metadata via OAI. We demonstrated this channel at the Internet2-DSI conference,

2002. We have assisted sister projects that are promoting learning in computing by making our technologies available.

## 4.5. To the public welfare beyond science and engineering?

We have promoted OAI, which is broadly supporting sharing of knowledge.