Annual Report 2002

Open Archives: Distributed Services for Physicists and Graduate Students OAD

to the Deutsche Forschungsgemeinschaft (DFG) and National Science Foundation (NSF)

Prof. Dr. Edward A. Fox,
Virginia Polytechnic Institute and State University, Dept. of Computer Science, Blacksburg, Virginia, USA
Dr. Heinrich Stamerjohanns, Prof. Dr. Eberhard R. Hilf,
Institute for Science Networking Oldenburg, Dept. of Physics, Carl von Ossietzky University, Oldenburg, Germany
Prof. Dr. Elmar Mittler,
Niedersächsische Staats- und Universitätsbibliothek Göttingen,
Prof. Dr. Royce Zia,
Virginia Polytechnic Institute and State University, Dept. of Physics, Blacksburg, Virginia, USA

1. Participants

1.1. What people have worked on the project?

Virginia Polytechnic Institute and State University, Blacksburg, USA

- Prof. Dr. Edward A. Fox
- Marcos A. Gonçalves
- Prof. Dr. Royce Zia
- Ye Zhou

Institute for Science Networking, Oldenburg, Germany

- Kerstin Zimmermann has worked for this project in January and February 2001, before she left for a job in Austria.
- Dr. Heinrich Stamerjohanns joined the project in March 2001.
- Prof. Dr. Eberhard R. Hilf
- Carsten Poppen and Svend Age Biehs work as students.
- Michael Schlenker, Diploma student

1.2. What other organizations have been involved as partners?

Official cooperation partners in the project are

- SUB Niedersächsische Staats-und Universitätsbibliothek Göttingen (Prof. Dr. E. Mittler)
- Computer Center RZ at HUB Alexander von Humboldt-Universität Berlin (HUB-OAi project: S. Dobratz)

1.3. Have you had other collaborators or contacts?

We are working in collaboration with the project *MathDiss International*, a DFG project at University of Duisburg (Prof. Dr. Törner) and have jointly organized a workshop in September 3, 2001, in Duisburg. We are working with the American Physical Society (APS), the Institute of Physics (IOP) Publishing, London, and have made contacts with other commercial publishers. From APS and IOP we have obtained data from all their publications for classification experiments. We support the Deutsche Initiative für Netzwerkinformation (DINI) by proposing and supporting national training workshops on the Open Archives Initiative. We collaborate with researchers at Federal University of Minas Gerais in Brazil: Dr. Berthier Ribeiro Neto, Dr. Alberto Laender, and Pavel Calado, and researchers at the Old Dominion University: Dr. Kurt Maly, Dr. Mohammad Zubair, and Xiaoming Liu. We collaborate with the staff of the CITIDEL (Computing and Information Technology Interactive Digital Educational Library) project based at Virginia Tech in classification and crawling experiments. There is also a general collaboration with various participants in the international Open Archives Initiative. By January 2002 a fruitful contact was opened with Dr. Thomas Krichel, Long Island University, NY, USA.

2. Activities and Findings

2.1. What were your major research and education activities?

The objective of the project is to improve the quality of resources and distributed digital library services, aimed at two communities: physicists and graduate students. The approach is to apply Open Archives Initiative (OAI) ideas and concepts to the physics community and the Networked Digital Library of Theses and Dissertations (NDLTD). For the two previous cited communities we are building a number of OAI-based digital library services and software tools, including:

- Classification services and tools based on PACS (Physics and Astronomy Classification Scheme) and APS/IOP data
- Crawlifier (crawler + classifier)
- Multi-ontology browsing services for APS/IOP articles and PhysNet (which is run by Professor Hilf)
- NDLTD Union Collection: searching and browsing services
- 5SLGen for MARIAN (developed at Virginia Tech)

- 5SGraph modeling tool
- PhysJob service
- Individuals NDLTD repository
- New NDLTD registration service
- The Web-DL environment.
- Open Digital Library and DL-in-a-Box concepts and prototype implementations

Since early 2001, we have made documents collected by the PhysDoc Harvest^{TM} system available to the Open Archives community through the OAI Protocol for Metadata Harvesting. The current heterogeneous collection of 40000 documents in PhysDoc, which is a collection of physics related documents throughout the world, has been examined for usability. The usefulness and quality of the automatically generated metadata from these documents has been examined and tested. We have focused our research on the development of mechanisms for the generation of suitable metadata converters of Dublin Core metadata and the generation of quality maintenance functions, so that documents with unusable metadata are not included in this OAI-collection.

In March 2001, the heterogeneous document archive PhysDoc was made OAI-compliant (see register at openarchives.org) and acts now as an OAI Data-Provider. Through the OAI-PMH interface, Virginia Tech collects the metadata and combines it into an Open Digital Library (ODL) network to provide DL services atop the union archive of metadata maintained by the Networked Digital Library of Theses and Dissertations (NDLTD).

Since June 2001, PhysDoc also has been running as an OAI Service Provider. It collects documents from many different sources and offers a search interface (see http://www.physdoc.org/query.php) to access those documents. This service provider has been integrated to the PhysDoc service of the European Physical Society (EPS) for better outreach. Since fall 2001, the metadata of articles of the publisher IOP has been included. This involved developing a data-mapper from their internal metadata format to Dublin Core. The metadata is automatically updated every night, after parsing XML data files that are produced by IOP and sent to us by email.

In Spring 2002, the conversion of the MARIAN system to Java was completed. The system now serves as a platform for a number of experiments related to this research project, as well as others undertaken in the Virginia Tech Digital Library Research Laboratory (DLRL). MARIAN will be used to test new algorithms combining belief networks, which have the potential to improve the quality of current PhysNet searching services. Further use of the MARIAN prototype system will be analyzed, since PhysNet must run in production mode.

Starting Fall 2002, we have been investigating classification tools and algorithms in order to support and improve a number of services for the Physics community including:

1) Automatic classification and flexible browsing of PhysNet and APS/IOP collections based on multiple ontologies
    a. This service could extend the current browsing capabilities of APS, IOP, and PhysNet – therefore offering multiple "organizational views" and access entry points to these collections, as well as enhancing the "resource discovery" capabilities of those sites.
2) Crawlifier
    a. The trained classifiers can be used to support focused crawling of the Web, therefore improving coverage beyond that of the current set of resources made available by PhysNet.

A number of classifier tools and algorithms have been tested and tuned, using data provided by APS, IOP, and PhysNet. Most of this data was harvested from OAI repositories. All data has been used with the consent of the collaborating partners. Classification results have been encouraging and the trained classifiers will now be applied and tested to support the proposed services. Prototype browsing and classification services have already been implemented and are now being packaged as ODL components to foster reuse and interoperability. Similar techniques are being applied to help the computing community in the CITIDEL project.

In December 2002, three prototypes were released, namely versions of the ODL component suite, 5SGraph, and 5SLGen. Those efforts bring us closer to the objective of supporting the whole process of automatic generation of tailored digital libraries, covering the entire development sequence, from requirements gathering to modeling to implementation. Current efforts are concentrated on integrating all of the tools and approaches into a coherent framework.

Regarding NDLTD, there were two major developments in late 2002. The first one was the development of a system based on the ODL components to register and update information regarding new and current members of the NDLTD. The old process was completely manual and very time-consuming. The new system allows new members to register themselves and old members to manage their own information. The second major development was the release of a new software environment, called Web-DL, which enables us to incorporate the content of NDLTD members that publicize their ETDs only through the Web. This is important in order to cover small universities and institutions that cannot afford or are not willing to maintain an Open Archive with their content. An experimental NDLTD Union Catalog created with the environment was roughly 30%

larger than the official Union Archive in terms of individual ETDs collected, and covered almost double the number of individual universities.

2.2. What are your major findings from these activities?

- Acquiring data, testing, and tuning classification algorithms are very collection-dependent and time-consuming tasks. In particular, working with commercial publishers was shown to be very difficult.
- Metadata quality and provenance are essential for building good services.
- Extracting information from the Web in a structured way is also a difficult endeavor. However, the work of generating DLs from the Web has been shown to have great potential in restricted, community-oriented domains.
- There is some difficulty in providing widespread OAI-based services due to some initial impedance with regard to the several archives involved in the project. For example, it is infeasible to impose a requirement upon all NDLTD members to adopt (in a short period of time) the Open Archives Protocol for Metadata Harvesting (OAI-PMH). We have developed toolkits and other software tools to help with this difficult task. Also, an alternative architecture involving Web-DL has been shown to provide a reasonable alternative in some cases.
- OAI services are sometimes hard to build; they are occasionally dependent on specific collection properties. For example, some PhysJob modules are dependent on the structure of the particular metadata standard we are using.
- The 5S approach and family of tools and languages has been shown to be very promising. It supports the tasks of rapid modeling, prototyping, and generation of digital libraries.
- The idea of Open Digital Libraries can be used to modularize some of the services we are building so that interoperability and flexibility can be achieved.
- The MARIAN system has state-of-the-art properties (e.g., semantic network structures, weighting schemes, and an object-oriented DL API) that can be very useful in building high-quality services. However, its monolithic architecture is very complex and makes it hard to deploy in some contexts. Also, more work is needed to ensure robustness in the current system.
- For MARIAN to be converted to a componentized system would require a partial rethinking of its architecture to achieve more

decoupling and easier application to new user groups and collections, as well as broader user adoption.

- It has been shown that the OAI protocol is suitable to include various heterogeneous sources into one union catalog in order to offer uniform access to such different archives as collections of grey literature, preprint-servers, and peer reviewed articles.
- Many users, especially students, interested in either educational or up-to-date scientific papers are not aware (and should not need to know) of the various publishers in physics. Through the OAI protocol it is fairly easy to interconnect very different sources and present them through one easy to use search interface, hiding unnecessary details.
- NDLTD needs a more formal organization and further automation. Initial deployment of ODL tools has allowed partial automation of a registry and related member services. There are plans for NDLTD to become a non-profit (501 c 3) organization.

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 as well as for the physics research community. A set of software tools was developed to give support to those services (see 3.3). Multiple courses at Virginia Tech (especially CS5604, Information Storage and Retrieval) have had one or more project groups learning through involvement in this effort. In June 2001, Heinrich Stamerjohanns gave training sessions at the University Library of Stuttgart (20 participants throughout Southern Germany) and at the Computer Center at Humboldt-University Berlin (30 participants from Northern Germany) to give an introduction to the Open Archives protocol and to present implementations of OAI Data and Service Providers. An introductory class on Information Retrieval and Digital Libraries also has been held at the University of Oldenburg. Tutorials and documentation on the ODL components have been developed and made available.

2.4. What outreach activities have you undertaken?

Many papers regarding the related efforts have been published in the major digital library and information retrieval conferences (JCDL, ECDL, ICADL, SIGIR, SPIRE, CIKM, etc. – see 3.1.2). Multiple tutorials have been given at digital library conferences by Edward A. Fox, Hussein Suleman, and Heinrich Stamerjohanns. To inform Europeans about the Open Archives Initiative, Heinrich Stamerjohanns gave a presentation at the European Open Archives Initiative Day on February 26, 2002 in Berlin. A workshop on Open Archives was jointly organized by 1) State Library of Lower Saxony and the University Library of Goettingen, Germany (partner in the OAD project) ; 2) Institute for Science Networking at the Carl von Ossietzky University, Oldenburg, Germany (partner in the OAD project); 3) Faculty of Science-Department of Mathematics, Gerhard-Merkator-University of Duisburg, Germany (DFG-Project *MathDiss International*); with the title *International Interdisciplinary Open Archives and Subject Specific Services in Mathematics and Physics*, September 3, 2001. There were speakers, topic lists, and links to online copies of the presented materials. Edward A. Fox and Marcos A. Gonçalves, at ECDL 2001 in Darmstadt, and at SIGIR 2001 in New Orleans, have given demonstrations about MARIAN. Prof. Edward A. Fox organized a workshop about Open Archives at ACM SIGIR'01 in New Orleans. There the current work of the project was presented by Edward A. Fox and Heinrich Stamerjohanns. The experimental search interface to the OAI-Service provider has been included in the Physdoc service of European Physical Society (EPS). For cooperative work, Marcos A. Gonçalves visited the Institute of Science Networking in Oldenburg in June, while Heinrich Stamerjohanns visited the group at Virginia Tech in September 2001. Prof. Fox visited Oldenburg in September, 2002, and an OAD meeting featuring both groups also was held during the ECDL 2002 conference in Rome. An ECDL 2002 panel on OAI componentized digital libraries also involved Dr. Fox.

3. Products

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

3.1.1. Major journal publications

1) Marcos André Gonçalves, Edward A. Fox, Layne T. Watson, and Neill A. Kipp. Streams, Structures, Spaces, Scenarios, Societies (5S): A Formal Model for Digital Libraries, 2003 (under review for *ACM Transactions on Information Systems*).

2) Qinwei Zhu, Marcos André Gonçalves, Edward A. Fox. 5SGraph: A Domain-Specific Visual Modeling Tool for Digital Librarians, 2003, now being submitted to the *Journal of the American Society for Information Science and Technology (JASIST).*

3) Edward A. Fox, Marcos André Gonçalves, Gail McMillan, John Eaton, Anthony Atkins, and Neill Kipp. The Networked Digital Library of Theses and Dissertations:

Changes in the University Community, *Journal of Computing in Higher Education*, vol. 13, number 2, pages 102-124, 2002.

4) Marcos André Gonçalves and Edward A. Fox, Technology and Research in a Global Networked University Digital Library, Marcos André Gonçalves and Edward A. Fox. *Revista Ciência da Informação* (Leading Information Science Journal in Brazil), vol. 30, no. 3, pages 13-23, 2001.

5) Hussein Suleman, Anthony Atkins, Marcos André Gonçalves, Robert K. France, Edward A. Fox, Vinod Chachra, Murray Crowder, Jeffrey Young. Networked Digital Library of Theses and Dissertations: Bridging the Gaps for Global Access - Part 2: Services and Research**,** Hussein Suleman, Anthony Atkins, Marcos André Gonçalves, Robert K. France, Edward A. Fox, Vinod Chachra, Murray Crowder, Jeffrey Young: *D-Lib Magazine 7(9)*, 2001.

6) Hussein Suleman, Anthony Atkins, Marcos André Gonçalves, Robert K. France, Edward A. Fox, Vinod Chachra, Murray Crowder, Jeffrey Young. Networked Digital Library of Theses and Dissertations: Bridging the Gaps for Global Access - Part 1: Mission and Progress, *D-Lib Magazine 7(9),* 2001.

7) Edward A. Fox, Fernando Adrian Das Neves, Robert France, Marcos Gonçalves, and Hussein Suleman. Short piece (In Brief) Digital Libraries 2000, Fifth ACM Conference on Digital Libraries, (2001), *D-Lib Magazine 7(9)*.

3.1.2 Conference/Workshop Proceedings

1) Marcos André Gonçalves, Edward A. Fox: 5SL: a language for declarative specification and generation of digital libraries. Proceedings of *Second ACM/IEEE Joint Conference on Digital Libraries*, pages 263-272, July 2002, Portland, Oregon, USA.

2) Marcos André Gonçalves, Ming Luo, Rao Shen, Mir Farooq Ali, Edward A. Fox: An XML Log Standard and Tool for Digital Library Logging Analysis, Proceedings of 6th European Conference on Research and Advanced Technology for Digital Libraries, pages 129-143, Rome, Italy, September 16-18, 2002

3) 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: Proceedings of the 9th International Symposium on String Processing and Information Retrieval, pages 194-209, Lisbon, Portugal, September 11-13, 2002.

4) Pável Calado, Altigran Soares da Silva, Berthier A. Ribeiro-Neto, Alberto H. F. Laender, Juliano P. Lage, Davi de Castro Reis, Pablo A. Roberto, Monique V. Vieira, Marcos André Gonçalves, Edward A. Fox: Web-DL: an experience in building digital libraries from the web. Proceedings of the 2002 ACM CIKM International Conference on Information and Knowledge Management, pages 675-677, McLean, VA, USA, November 4-9, 2002.

5) Marcos André Gonçalves, Robert K. France, Edward A. Fox. MARIAN: Flexible Interoperability for Federated Digital Libraries, *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries,* pages 172-186, Darmstadt, Germany, September 4-9, 2001.

6) Marcos André Gonçalves, Ali A. Zafer, Naren Ramakrishnan, Edward A. Fox. Modeling and Building Personalized Digital Libraries with PIPE and 5SL, *Proceedings of the Joint DELOS-NSF Workshop on Personalization and Recommender Systems in Digital Libraries,* Dublin, Ireland, June 1-2, 2001.

7) Edward A. Fox, Robert France, Marcos André Gonçalves, Hussein Suleman**,** Building Interoperable Digital Library Services: MARIAN, Open Archives, and NDLTD, *Proceedings of The 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 451, New Orleans, Louisiana, September 9-13, 2001.

8) Marcos André Gonçalves, Robert K. France, Edward A. Fox, Eberhard R. Hilf, Michael Hohlfeld, Kerstin Zimmermann, Thomas Severiens, Flexible Interoperability in a Federated Digital Library of Theses and Dissertations, *Proceedings of 20th World Conference on Open Learning and Distance Education,* Dusseldorf, Germany, 01-05 April 2001.

9) Marcos André Gonçalves, Robert K. France, Edward A. Fox, Tamas E. Doszkocs, MARIAN Searching and Querying of Heterogeneous Federated Digital Libraries, *Proceedings of First DELOS workshop on "Information Seeking, Searching and Querying in Digital Libraries",* Zurich, Switzerland, September 4-9, 2000.

10) Heinrich Stamerjohanns, *Implementing OAI Data and Service Providers*, in Proceedings of the International Interdisciplinary Open Archives and Subject Specific Services in Mathematics and Physics, Duisburg, 2001

3.1.3. Books and other one-time publications

1) The UNESCO Guide (http://etdguide.org) is online available. The ETD Sourcebook will be published soon.

2) Edward A. Fox, Gail McMillan, Hussein Suleman, Marcos André Gonçalves, Networked Digital Library of Theses and Dissertations, chapter in *Digital Libraries: Policy, Planning and Practice*, eds. Derek Law and Judith Andrews, Ashgate Publishing, UK, scheduled for 2003

3) Edward A. Fox, Marcos A. Gonçalves and Neill A. Kipp, Digital Libraries, *In Handbook on Information Technologies for Education & Training, in Springer series "International Handbook on Information Systems"*, ed. Heimo Adelsberger, Betty Collis, Jan Pawlowski, 2001, 19 pages, (2001)

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

The project page can be found under http://www.physdoc.org/OAD.html. It provides information about the personnel of both groups at Oldenburg and at Virginia Tech, related institutions, and present services, namely MARIAN and PhysDoc. More information about MARIAN also can be found directly at http://www.dlib.vt.edu/projects/MarianJava/index.html. Information about NUDL can be found at http://www.nudl.org.
Other pages include

- Physics jobs http://physjob.nudl.org
- 5S page at VT: http://www.dlib.vt.edu/projects/5S-Model/index.html
- OAI page at VT: http://www.dlib.vt.edu/projects/OAI/index.html
- ODL page at VT: http://oai.dlib.vt.edu/odl/

General information about the Open Archives Initiative can be found at http://www.openarchives.org, which documents the registry of the OAD add-on to PhysDoc as a Data Provider. Further information about Open Archives activities in Germany can be found at http://www.dini.de/dinioai/dinioai.php3.

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 NDLTD Union Catalog, searching services

The NDLTD Union Archive periodically harvests ETD metadata from NDLTD members that implement the OAI-PMH (NDLTD data providers) using software developed in the context of this and related projects. Search services were built to allow searching across the unified catalog, and extended services are currently in development.

3.3.2 Java MARIAN

MARIAN is a digital library system designed and built to store, search, retrieve, and browse large numbers of diverse objects in a network of relationships. MARIAN is built upon three basic principles: 1) unified representation based on *semantic networks*, which model internal structures of digital objects and metadata and different types of relationships among objects and concepts (e.g., as in thesauri, classification hierarchies or among word terms and structural portions of documents); 2) *weighting schemes* to support information retrieval services, including weighted nodes and links, and weighted objects sets wherein a set of objects whose relationship to some external proposition is encoded in their decreasing weight within the set; and 3) an *object-oriented class system*, which is used to organize nodes and links into hierarchies of object-oriented classes, with methods to store and maintain instance objects of their class, translate back and forth between object IDs and fully realized objects, and support matching and retrieval operations.

MARIAN has been completely re-engineered and re-implemented in Java and is now one of the main research tools of the DLRL; new and existing projects of the DLRL make use of MARIAN due to its modular and extensible architecture, powerful and flexible representation model, and other invaluable characteristics.

3.3.3 5SLGen for MARIAN

A digital library generator was developed which can take 5SL DL specifications and generate an implementation of a DL directed towards the MARIAN DL system. It is now being tested, using the MARIAN DL system, to provide prototype DL services across the NDLTD Union catalog, CITIDEL, the PhysNet collection, and other applications.

3.3.4 PhysJob service

PhysJob is a service that aims to serve academia and high school teachers by publishing resumes and job opportunities posted by physics departments and high school principals. It allows administrative editorial control to ensure the accuracy of the database. It can use OAI MPH to harvest from job information providers.

3.3.5 5SGraph modeling tool

5SGraph is a domain-specific visual tool aimed at modeling digital libraries. 5SGraph is based on a metamodel that describes digital libraries in a high-level based on a formal theory -- the 5S theory. The output from 5SGraph is a DL model that is an instance of the metamodel, expressed in a digital library description language -- 5SL, which enables further automatic generation of digital libraries. 5SGraph enables component reuse to reduce the time and efforts of designers. Furthermore, 5SGraph maintains semantic constraints specified by the 5S metamodel and enforces these constraints over the instance model to ensure semantic consistency and correctness.

3.3.6 PACS browsing service

The PACS browsing tool allows hierarchical browsing of physics union catalog collections using the Physics & Astronomy Classification Scheme (PACS). Users may find all the documents under certain categories sorted by field through browsing and can jump to a closely related topic.

Currently, the APS (American Physics Society) online journal collection is loaded and the IOP (Institute of Physics) electronic journal collection is about to be imported.

3.3.7 Individuals NDLTD repository

An E-Prints server was installed and configured with ETD-MS, the metadata standard for electronic theses and dissertations, to allow individuals whose institutions are still not members of NDLTD to deposit their our own thesis or dissertation and allow us to make it available, after suitable administrative control is exercised.

3.3.8 Java Fulltext Document Object

This transforms PDF and PS documents into plain text format. It may help with construction of other services (e.g., cross-reference linking, full-document search, etc.).

3.3.9 Converter

The Dublin Core (DC) Metadata Standard lacks clear definitions of the representations to be used for the typical Dublin Core elements. As a result, many different descriptions for date and language identifiers are in use. We have developed several algorithms in order to normalize the descriptions to *best practice* representations, so, e.g., many different date representations are converted to ISO 8601, which is the recommended *best practice* representation for a date within DC. Several normalizers have been developed for the various DC elements. Since the Harvest project uses SOIF Metadata, a converter from SOIF to DC Metadata has been developed.

3.3.10 ODL components

Open Digital Libraries (ODLs) are systems built as networks of extended Open Archives. The basic philosophy adopts the notions of simplicity and reusability from the Open Archives Initiative, and adds extensibility and componentization into the mix. Protocols for inter-component communication with a single digital library are designed as extensions of the OAI Protocol for Metadata Harvesting, and then components that adhere to these protocols are composed to operate as the back-end of a DL.

3.3.10 NDLTD registration service

A new service has been developed which allows new NDLTD members to register themselves with the federation and automatically includes their information in the NDLTD databases and web pages, therefore substituting the old manual process of registration. Information about old members can be updated with the system too. The changes are automatically reflected in the NDLTD web pages. The service is built using the ODL components, therefore serving as an illustration of the reusability of the components.

3.3.12 Web structure crawler and GetSmart concept map generator

Web structure analysis can discover the underlying ontology of the web site and, furthermore, the structure itself can be regarded as a classification scheme. For example, in PhysNet, there are categories such as physics department, PhysDoc, conference, journal, etc. Therefore, in addition to the web page crawling, a web structure spider also can reveal the hierarchy of a specific website and generate an XML description of the web ontology. To better understand the data through data visualization, a GetSmart concept map generator has been provided to improve the usability. (GetSmart is a tool developed through collaboration of Virginia Tech with the University of Arizona, which is carrying out research in a project funded by NSF through the NSDL program.)

3.3.13 Physics document automatic classification service

The task to categorize a new document is always difficult for the author, especially against a complicated classification scheme like PACS. To help physicists and to expedite the process of finding the correct PACS code(s), a new automatic classification service has been created, which helps to identify the proper PACS code(s). This can work with other ontologies with an offline trained classifier. The classifiers will facilitate PACS-based navigation of resources not classified, e.g., those Web resources collected at PhysNet as well as a more focused crawling of physics resources in the Web.

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, light-weight protocol. We have demonstrated, in a small scale, the applicability of such concepts to build high quality services in cross-institutional and discipline levels.

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: Physics, Computer Science, and medical information (in conjunction with NLM/ORISE support.)

4.3. To the development of human resources?

We have introduced OAI to large segments of the Virginia Tech campus, and to many others in conjunction with our NSDL (www.nsdl.org) related and other efforts. We have involved students in multiple classes at Virginia Tech, who now have knowledge of these concepts and technologies. We have involved roughly 20 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. We have developed PhysJob to support physics teachers and researchers in their job-seeking efforts. As already mentioned in 2.3, two training workshops have been held in Germany especially for librarians in order to inform them about the Open Archives Initiative.

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

We have developed a union service for electronic theses and dissertations that is of broad interest. Content now includes the PhysDis resources, helping disseminate physics research more broadly. We have assisted sister projects that are promoting learning in computing by making our technologies available, including for the many projects related to NSDL.

4.5. To the public welfare beyond science and engineering?

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