Annual Report 2001 - 2002

Open Archives: Distributed Services for Physicists and Graduate Students (OAD)

to the Deutsche Forschungsgemeinschaft (DFG) and National Science Foundation (NSF).
A PDF Version of this document is also available.

Prof. Dr. Edward A. Fox,
Virginia Polytechnic Institute and State University (Virginia Tech), 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
- Robert K. France
- Ni Liu

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 has 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.
We support the Deutsche Initiative für Netzwerkinformation (DINI) by proposing and supporting national training workshops on the Open Archives Initiative.
We collaborate with other researchers at Virginia Tech: Hussein Suleman, Paul Mather, Ryan Richardson, Jun Wang, Ming Luo, Priya Shivakumar.
There is also a general collaboration with various participants in the international Open Archives Initiative.
We also started contacts to incorporate metadata from commercial publishers and have contacts with IOP Publishing, London.
Further interesting contacts include Dr. Kurt Maly and Xiaoming Liu, Old Dominion University, and the American Physical Society (APS).
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 scope of the project is to improve distributed digital library services, aimed at two communities: Physicists and graduate students. We try to apply Open Archives ideas and concepts to the physics community and the Networked Digital Library of Theses and Dissertations (NDLTD, see http://www.ndltd.org). For the two previous cited communities we are building OAI-based digital library services and software tools:

- NDLTD Union Collection and Searching Service
- 5SL DL generator
- PhysJob service
- PACS browsing service
- Individuals NDLTD repository
- Open Digital Library concept and prototype implementations

We also investigate the automated generation of Digital Libraries (DLs) from specifications. We have made documents collected by the PhysDoc Harvester available to the Open Archives community through the OAI Metadata Harvesting Protocol. See Figure 1.

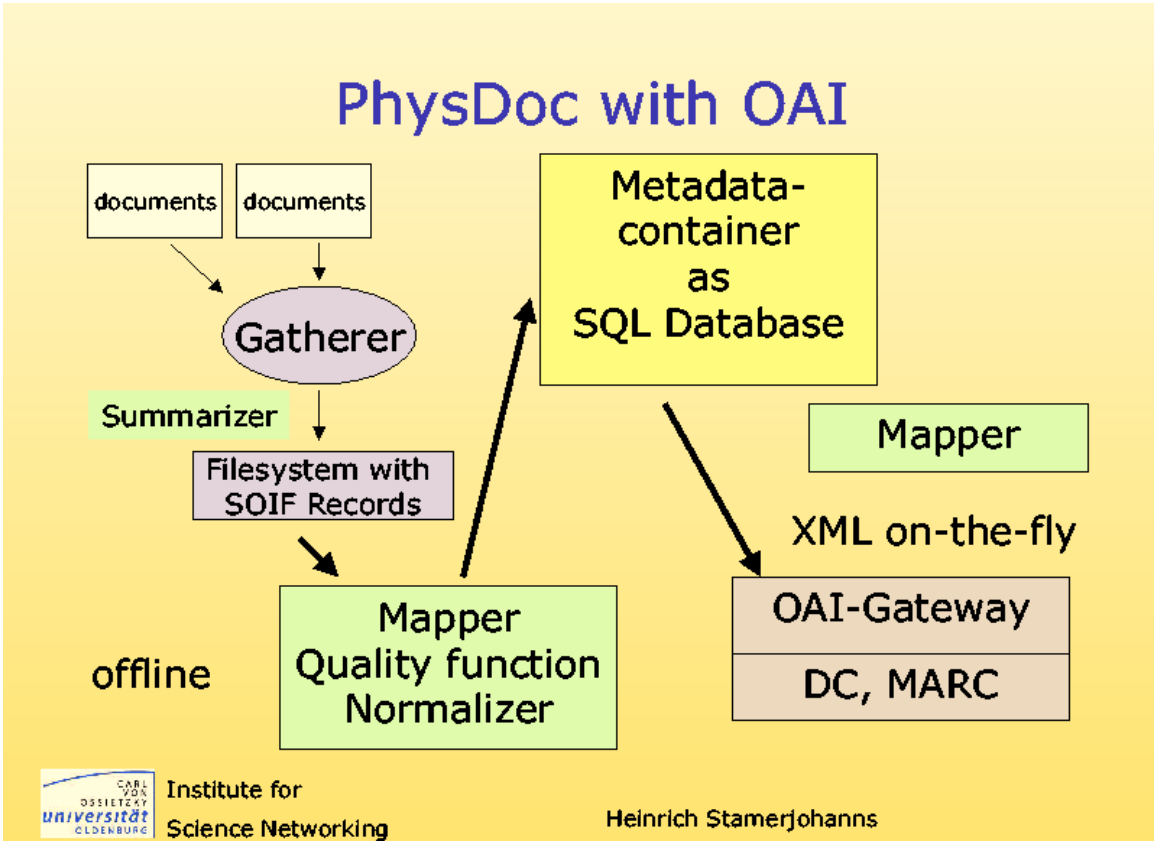

Figure 1. PhysDoc with OAI

The current heterogenous collection of 40000 documents in PhysDoc, which is a collection of physics related documents throughout the world, has been examined for usability in well-formed metadata collections. 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 the generation of suitable metadata converters of Dublin Core metadata and the generation of quality functions, so documents with unusable metadata are not included in this OAI-collection.
In March 2001, the heterogenous document archive PhysDoc was made OAI-compliant (see the registry at openarchives.org) and acts now as an OAI Data-Provider. Through the OAI-PMH interface, Virginia Tech collects the metadata combined into an Open Digital Library (ODL, see http://oai.dlib.vt.edu/odl/) network to provide DL services over 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 a OAI Service Provider (see Figure 2), by collecting documents from many different sources and offering a search interface (see http://www.physdoc.org/query.php) to access those documents. The Service Provider has been integrated into the PhysDoc service of the European Physical Society (EPS) for better outreach. Since fall 2001, the metadata of articles of the publisher IOP have been included by developing a data-mapper from their internal metadata to Dublin Core. The metadata is automatically updated every night, by parsing XML-datafiles, which are produced by IOP and sent to us by email.

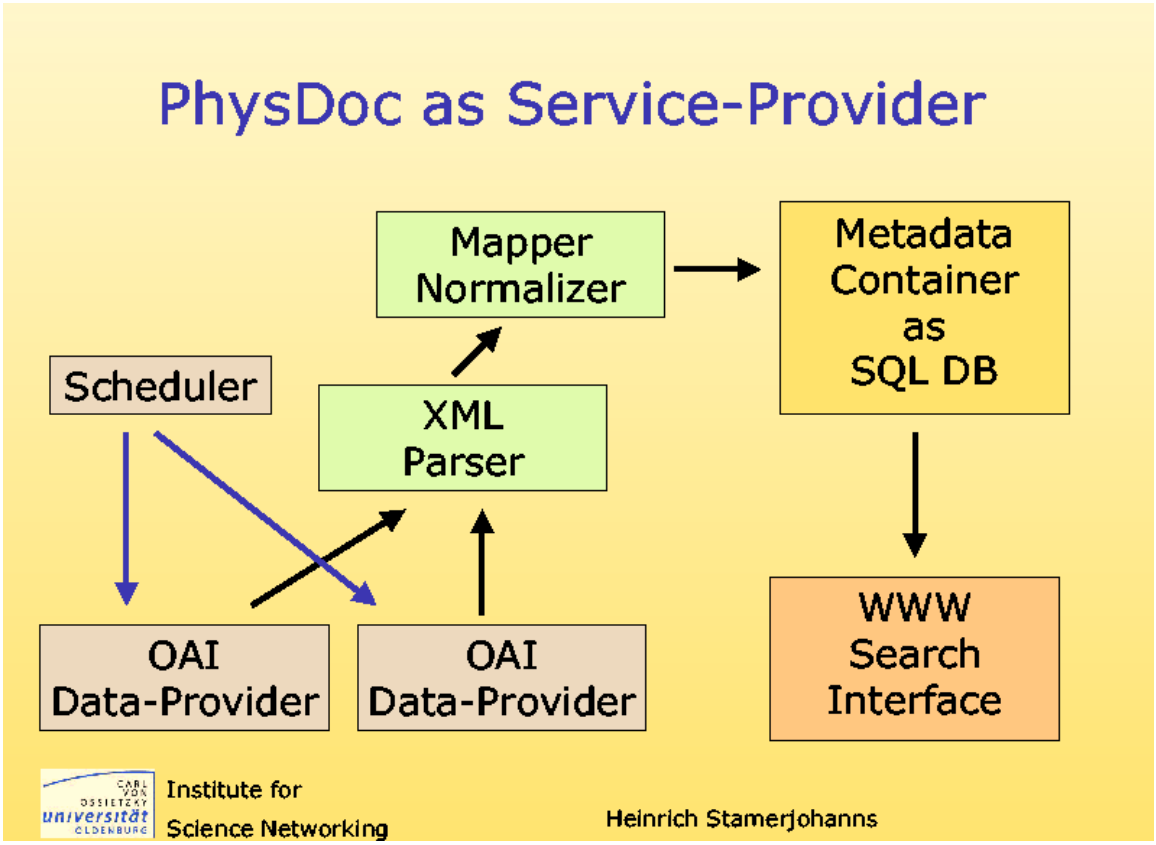

Figure 2. PhysDoc as Service-Provider

2.2. What are your major findings from these activities?

- Metadata quality and provenance are essential for building good services.
- There is some difficulty in providing widespread OAI-based services due to some initial impedance regarding the several archives involved in the project. For example, it would be impossible to make all the NDLTD members 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.
- 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 Language shows promise as a digital library requirements engineering tool and for automatic DL generation.
- The idea of an Open Digital Library can be used to modularize some of the services we are building, e.g., job service so that interoperability & flexibility can be achieved.
- 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 OAD-based services.
- The segmentation of MARIAN requires 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 heterogenous sources into one Union catalog in order to offer uniform access to such different archives: heterogenous collections of grey literature, preprint-servers, and peer reviewed articles.

- Learners, especially students, interested in either educational or up-to-date scientific papers generally are not aware (and should not need to know) 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.

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 in the Physics research communitity. A set of software tools was developed to give support to those services (see "What other specific products have you developed?"). 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 has been held at the University of Oldenburg.

2.4. What outreach activities have you undertaken?

Multiple tutorials have been given at digital library conferences by Edward A. Fox and 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, February 26, 2001, in Berlin.

A workshop on Open Archives was jointly organized by

- State Library of Lower Saxony and the University Library of Goettingen, Germany (partner in the OAD project)
- Institute for Science Networking at the Carl von Ossietzky University Oldenburg, Germany (partner in the OAD project)
- 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. The workshop site identifies speakers, topics, and links to the talks and presented materials.

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

Prof. Edward A. Fox has organized a workshop about Open Archives at ACM SIGIR’01 in New Orleans. There the current work of the project has been presented by Edward A. Fox and Heinrich Stamerjohanns.

The experimental search interface to the OAI-Service provider has been included into the Physdoc service of European Physical Society (EPS). For cooperative work, Marcos A. Gonçalves has visited the Institute of Science Networking in Oldenburg in June, while Heinrich Stamerjohanns has visited the group at Virginia Tech in September.

3. Products

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

3.1.1. Major journal publications

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.* In special issue on "Information Technology and Educational Change" of The Journal of Computing in Higher Education, 13(2): 3-24, Spring 2002, in press.

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.* In revision for possible publication in ACM Transactions on Information Systems (TOIS), 2002.

Marcos André Gonçalves and Edward A. Fox, *Technology and Research in a Global Networked University Digital Library (NUDL).* In Ciência da Informação, 30(3):13-23, December 2001.

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

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

3.1.1.1 Conference/Workshop Proceedings

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, Düsseldorf, Germany, 01-05 April 2001

Marcos André Gonçalves, Robert K. France, Edward A. Fox, *MARIAN: Flexible Interoperability for Federated Digital Libraries.* In Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL2001), Darmstadt, Germany, September 2001.

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

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

3.1.2. Books and other one-time publications

The UNESCO Guide (http://etdguide.org) is online available.

The ETD Sourcebook is in process for publication by Marcel Dekker (of New York).

Edward A. Fox, Marcos A. Gonçalves and Neill A. Kipp. *Digital Libraries. In Handbook on Information Technologies for Education & Training*, Springer series "International Handbook on Information Systems", ed. Heimo Adelsberger, Betty Collis, Jan Pawlowski, 2002, 19 pages. (Book Chapter)

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 the personnel of both groups at Oldenburg and at Virginia Tech, its institutions, and its present services: MARIAN and PhysDoc. More information about MARIAN 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/ 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/, 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 and search 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 5SL DL generator

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

3.3.2 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 keep the accuracy of the database. It can use OAI-PMH to harvest from job information providers.

3.3.3 PACS browsing service

The PACS browsing tool allows hierarchical browsing of Physics collections using the Physics & Astronomy Classification Scheme (PACS). As different disciplines have their own classification systems (e.g., ACM classification system for computing), we built a generic generation tool for classification scheme browsing that takes a canonical XML Schema for the classification and communicates with the MARIAN digital library system.

3.3.5 Individuals NDLTD repository

An E-Prints server was installed and configured with the ETDMS metadata standard 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.

3.3.6 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.)

3.3.7 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 dates and other 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 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.

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 NSDL (www.nsdl.nsf.gov) 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 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, ultimately for NSDL.

4.5. To the public welfare beyond science and engineering?

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