Annual Report 2003

Open Archives: Distributed Services for Physicists and Graduate Students OAD

to the Deutsche Forschungsgemeinschaft (DFG) and National Science Foundation (NSF). This report only covers the year 2003. Please see our previous report for the years 2001 and 2002.

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, master student
- Yuxin Chen, Ph. D. student
- Baoping Zhang, Ph. D. Student

Institute for Science Networking, Oldenburg, Germany

- Dr. Heinrich Stamerjohanns
- Prof. Dr. Eberhard R. Hilf
- Andreas Werner, Diploma student
- Eike Bernhardt, Diploma student
- Michael Schlenker, Ph. D. 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, Dr. H. Neuroth)

- Computer Center RZ at HUB Alexander von Humboldt-Universität Berlin (HUB-OAi project: S. Dobratz, U. Müller)

1.3. Have you had other collaborators or contacts?

We support the Deutsche Initiative für Netzwerkinformation (German Initiative for Networked Information, DINI) by proposing and supporting national training workshops on the Open Archives Initiative. 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. We also collaborate with Dr. Thomas Krichel, Long Island University, NY, USA and Prof. Dr. Stevan Harnard, University of Montreal.

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) like SHRIMPS,
- Multi-ontology browsing services for APS/IOP articles and PhysNet (which is run by the European Physical Society (EPS) in cooperation with the ISN, Oldenburg)
- NDLTD Union Collection: searching and browsing services
- 5SLGen for MARIAN (developed at Virginia Tech)
- 5SGraph modeling tool
- PhysJob service
- Individuals NDLTD repository
- NDLTD registration service
- The Web-DL environment.
- Open Digital Library and DL-in-a-Box concepts and prototype implementations
- Physics filtering service from NDLTD

The documents collected by the PhysDoc Harvest system (self-archived articles througout worldwide physics departments) has been expanded to approx. 100,000 articles by the use of the crawlifier SHRIMPS.

SHRIMPS will automatically find physics relevant scientific document pages on web servers of all the worldwide distributed institutions listed in PhysDep. The SHRIMPS system is being adapted and used for enriching the searchable PhysDoc index. Further countries besides Germany are processed by the system and will be included step by step into the service. Recently, Finland has been added, with Croatia, Sweden and most other European countries due for inclusion soon.

This heterogeneous collection of 100,000 documents is being made available to the PhysDoc OAI-Service Provider by suitable metadata converters.
The amount of collected documents by the PhysDoc OAI Service Provider has been expanded to 410,000 documents. We have tried various fulltext index systems and have now switched to the *tsearch2* engine of the postgresql database.
The PhysDoc OAI Service Provider collects documents from many different sources and offers a search interface (see http://www.physnet.net/query.php) to access those documents. The quality of the collected metadata has improved in the last years, especially with the help of the OAI Metadata Harvesting Protocol. Formal standards are now more closely followed, but the metadata content has still to be improved. In order to improve the quality of the collected metadata, we have written a tool to check the quality of Dublin Core offered by various OAI Data Providers. The DC Checker is available at http://www.physnet.net/oad/dc.

2.2 What are your major findings from these activities?

- 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.

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. The set of software tools has been further developed and expanded 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. H. Stamerjohanns has given tutorials on OAI at 3rd Workshop of the OAForum, March 27-29th, Berlin, 2003 (available online at http://www.dini.de/documents/oaforum3_tutorial_de.pdf) and at the ETD 2003 in Berlin 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.

The experimental search interface to the OAI-Service provider has been included in the Physdoc service of European Physical Society (EPS).
For cooperative work, Edward A. Fox visited Oldenburg in September 2003 in order to attend the ISN Summer meeting, where he presented current developments.

3. Products

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

3.1.1. Major journal publications

1) Baoping Zhang, Marcos Andre Goncalves, Yuxin Chen, Edward A. Fox, Pavel Pereira Calado, *Combining support vector machines and structural rules for effective filtering of OAI-based repositories, s*ubmitted to Journal of Digital Libraries.

3.1.2 Conference/Workshop Proceedings

1) Heinrich Stamerjohanns, *Lessons learned implementing Service Providers*, Workshop at JCDL 2003, May 27-29th, Houston.
2) Heinrich Stamerjohanns, *Tutorial on Open Archives*, Tutorial given at ETD2003, May 24-25th, Berlin, 2003.
3) Heinrich Stamerjohanns, *Tutorial on Open Archives*, Tutorial given at 3rd Workshop of the OAForum, March 27-29th, Berlin, 2003.
4) Heinrich Stamerjohanns, *Interoperability with Open Archives*, IuK 2003, Sharing Knowledge: Scientific Communication, March 10-13th 2003, Osnabrück.

3.1.3. Books and other one-time publications

1) The UNESCO Guide (http://etdguide.org) is online available. The ETD Sourcebook has been published 2003.

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, 2003.

3) Arbeitsgruppe Open Archives Initiative in Deutschland. *Elektronisches Publizieren an Hochschulen: Inhaltliche Gestaltung der OAI-Schnittstelle – Empfehlungen*: DINI Dokumente des DINI Servers, 2003.

4) Arbeitsgruppe Open Archives Initiative in Deutschland. *Electronic Publishing in Higher Education: How to design OAI interfaces - Recommendations* : DINI Dokumente des DINI Servers, 2003.

5) Edward A. Fox, *Research Challenges to Semi-Automatically Enhance Quality in Distributed Open Archives,* SINN03 eProceedings, Proceedings of the conference on Worldwide Coherent Workforce, Satisfied Users- New Services For Scientfic Information - September 17 - 19 2003, Oldenburg, Germany.

6) Ye Zhou, *Reengineering PhysNet in the uPortal framework,* Master Thesis, ETDs @VT, http://scholar.lib.vt.edu/theses/available/etd-06092003-164230/

7) Kevin Meyers, Stephen Foret, *Course Project(PhysDep) Report for VT CS6604 Digital Libraries,* http://tuppence.dlib.vt.edu/~yuchen/PhysDep/DLProjPhysnet.pdf

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

The project page can be found under http://www.physnet.uni-oldenburg.de/oad/. It provides information about the personnel of both groups at Oldenburg and at Virginia Tech, related institutions, and present services, namely MARIAN and PhysDoc.

- Physics jobs http://physjob.nudl.org
- 5S page at VT: http://www.dlib.vt.edu/projects/5S-Model/index.html
- OAD page at VT: http://www.dlib.vt.edu/OAD
- 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/dini/arbeitsgruppe/arbeitsgruppe_details.php?ID=9.

The project is also reflected and mentioned on the following sites:

- http://www.ub.uni-duisburg.de/mathdiss/work2.html
- http://www2.sub.uni-goettingen.de/cgi-bin/ssgfi/anzeige.pl?db=meta&nr=000555&ew=SSGFI
- http://www.inforum.cz/inforum2003/english/prispevek.asp?CisloSekce=8&Kod=7
- http://grp.lib.msu.edu/diglibdocumentation.html
- http://www.oaforum.org/otherfiles/oaf_d43_workshop2.pdf

- http://w210.ub.uni-tuebingen.de/dbt/volltexte/2003/826/html/festplatte/rundgang/pdf/nuessle.pdf
- http://webdoc.sub.gwdg.de/edoc/p/fundus/5/brahms.pdf
- http://www.minervaeurope.org/publications/globalreporthtm/germany.htm

3.3. What other specific products have you developed?

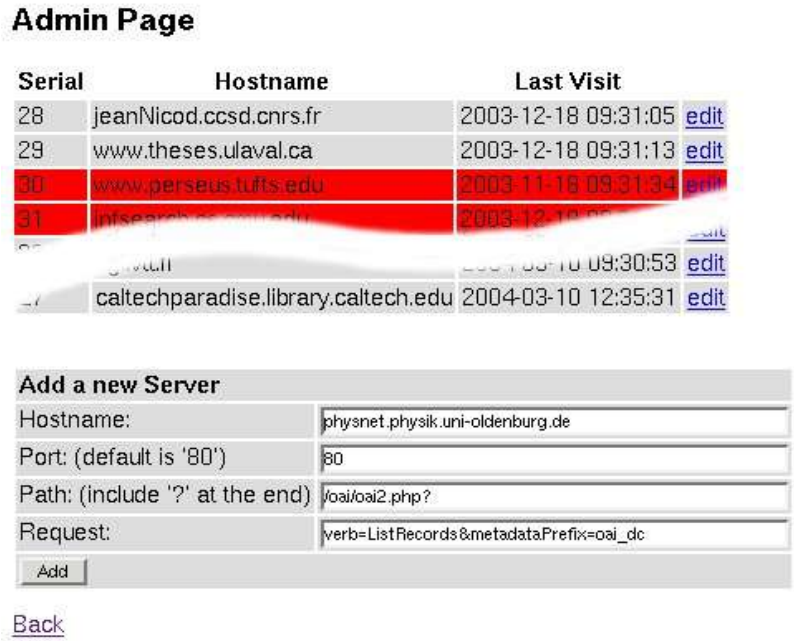
In additon to the products we have mentioned in our last report, we have built several software tools and packages for development of OAI-based services, including:

3.3.1 OAI Data-Provider

Additionally, an OAI-Client, which can be easily configured and supports multiple databases has been developed in order to ease the installation of OAI – based services. This client is Open Source (http://physnet.physik.uni-oldenburg.de/oai/), and has been installed
in several university libraries in Germany, and other countries like Belgium, Italy, India, Japan, Spain and USA.

3.3.2 OAI Service-Provider

An OAI Service Provider collects information from various OAI Data Providers. Our Service Provider collects only physics related documents from various sources and presents them through a unified search interface. We have expanded our collection to about 410.000 physics articles.



With a simple Administraor GUI additional sources can be easily added to our collection of sources:

The list of available sources can be also easily edited:

Editing Server

Repository Name:	Physnet, Physics Document Server, Germany
Admin Email:	mailto:stamer@uni-oldenburg.de
Hostname:	physnet.uni-oldenburg.de
Port: (default is '80')	80
Path: (include '?' at the End)	/oai/oai2.php?
Request:	verb=ListRecords&metadataPrefix=oai_dc
Granularity:	YYYY-MM-DD
Compression(s):	gzip
Friends:	
Query:	Yes / No
Reason:	
	Change

The implementation is Open Source and is available at http://physnet.uni-oldenburg.de/oad.

3.3.3 DC-Checker

The DC-Checker checks the validity of Dublin Core metadata. It does not focus on formal XML correctness of OAI-Records (the Repository Explorer of VT takes care of that) but checks whether the given Dublin Core Metadata follows the recommendations of the DCMI (plus additional checks).

Check your records with DC Checker.

These are total statistics taken from the number of hosts below:
Total hosts: 127

Connection errors: 55 — 43.31%
Hosts ok: 72 — 56.69%

Hosts with (supposedly) namespace errors: 1

dc:title

Total records: 37651

utf8 errors: 7 — 0.02%
empty records: 641 — 1.7%
dc errors: 0 — 0%
dc warnings: 83 — 0.22%
correct records: 36927 — 98.08%

dc:creator

Total records: 58235

3.3.4 SHRIMPS

SHRIMPS is an acronym for *Simple Http Robot to Improve the Physnet Services*. It intends to crawl web pages of physics institutions, following links, semantically*TTP Robot to Improve the Physnet Services*. It intends to crawl web pages of physics institutions, following links, analyzing the page titles and the first few paragraphs of pages it encounters, in order to find scientific document pages in physics and especially self-archived articles.
First tests on the University of Oldenburg physics departments web pages have shown a significant increase of about 70% percent in relevant objects in PhysDoc search repository when using SHRIMPS. The number of relevant objects in PhysDoc could be increased from 158 to over 270 for the Oldenburg department servers alone. Further experiments on the German part of PhysNet were conducted and similar improvements were found for German university sites.
This experiences led to the development and partial rewrite of the software into a distributed system of agents with a central server in order to provide a scaleable and performant solution for searching larger sets of institutions. Today, SHRIMPS consist of two parts, one is the workhorse of the system, a small program designed to be easily deployable on a wide range of hosts, called shrimps-agent, and a server program, which provides job management facilities and the web based user interface for the operators. The shrimps-agent queries web servers and decides which links to follow, while the shrimps-server aggregates the results, distributes the jobs under the agents and presents them to the operator in a human predictable way. The communication between the two program parts is based on W3C standard SOAP XML messaging technology (*http://www.w3.org/2000/xp/Group/*). To review and refine the results of the SHRIMPS system some additional tools have been developed, based on the Netscape/Mozilla web browser, allowing fast and easy review of the results produced by the SHRIMPS system.

3.3.5 OWL Ontology Web

To represent the structure of physics institutions and document repositories within a database adequately, we have developed an OWL Ontology Web Language (*www.w3c.org/2001/sw/WebOnt/*) semantic network. All information available directly or implicitly within the semantics and syntax of the HTML-files of PhysNet were transferred into the OWL network, represented as RDF triples within an SQL database.
This structure allows a deep structure of administrational data, which can easily be imported from other existing databases. By this, PhysNet will be enriched with information like postal addresses, contact information etc.
The OWL Ontology Web for PhysNet is currently under development, results and an implementation will be available by summer 2004.

3.3.6 Physics filtering service

Physics filtering service is used to automatically extract physics-related theses and dissertations from NDLTD union collection. The extracted physics sub-collection can be added into PhysNet to maximize the number of physics-related resources available to the physics community. This service combines evidences from support vector machine (SVMs) classifiers and structural rules (subject filter and contributor filter) to extract physics related theses and dissertations from the NDLTD collection. A belief network model was designed to formally support the combination process. The belief network

model provides a flexible and formal basis for the combination of several evidences, and can be applied to many other academic disciplines covered by NDLTD.

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.)

The provided software for data and service provider is used at various insitutes throughout the world.
The software for the e.g. data provider is used at the following repositories:

- Indiana University Digital Library Program
- ibiblio Software Collection
- Mormons and their Neighbors
- Library of Ruhr-Universitaet Bochum, GERMANY, Dissertation Server
- The Fitzwilliam Museum
- Collection FRAC aquitaine, panorama de l'art d'aujourd' hui
- ViFaPhys, The Physics Virtual Library
- Serveur des publications de l'Institut Francais d'Etudes Andines
- Virtual Library of Psychology at **Saarland** University and State Library, Germany, PsyDok
- GSI OAI Repository

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. We have offered various tutorials which helped people to create their own OAI compatible repositories and to get informed about the Open Archives Initiative. See also section 4.2.

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
Our results and implemented software has been included to PhysNet, which is a information service for Physicist and graduate students.

4.5. To the public welfare beyond science and engineering?

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