Workshop 1

ECDL 2000

Fourth European Conference on Research and Advanced Technology for Digital Libraries

18 – 20 September 2000 Lisbon / Portugal

Extending Interoperability of Digital Libraries:

Building on the Open Archives Initiative

21 September 2000

Lisbon, Portugal

Contents

Program

8:30 am – 9:00 am	**Registration**
9:00 am – 10:30 am	**Session One – Introduction to the Open Archives Initiative**
Introductory Remarks *Edward Fox, Virginia Tech* and *Carl Lagoze, Cornell University* (15 minutes) **Introductions from Workshop Participants** chaired by: *Edward Fox, Virginia Tech* and *Carl Lagoze, Cornell University* (30 minutes) A short introduction from each participant to familiarize people with one another and contextualize further interaction during the workshop. **Historical Overview** *Edward Fox, Virginia Tech* (45 minutes) Origins of the OAI and overview of the original technical agreements.	
10:30 am – 11:00 am	**Break**
11:00 am – 12:30 pm	**Session Two – Technical Details**
Expanding the Scope and New Technical Agreements *Carl Lagoze, Cornell University* (60 minutes) **Framing the Discussion for the Afternoon** *Edward Fox, Virginia Tech* (30 minutes)	
12:30 pm – 2:00 pm	**Lunch**
2:00 pm – 3:30 pm	**Session Three - Discussion**
Research Opportunities - Views of the Funders *Funding Agencies/Sponsors* (30 minutes) **General Discussion** chaired by: *Edward Fox, Virginia Tech* and *Carl Lagoze, Cornell University* (60 minutes) Focus on reactions to OAI agreements and their application to communities represented at the meeting.	
3:30 pm – 3:50 pm	**Break**
3:50 pm – 4:20 pm	**Session Four – Presentations**
Research Project Presentations *Constantino Thanos (IEI-CNR, Italy), Robert Tansley (University of Southampton, UK),* and *Eberhard Hilf (Oldenburg, Germany)*	
4:20 pm – 5:00 pm	**Session Five – Moving Forward**
Moving Forward chaired by: *Edward Fox, Virginia Tech* and *Carl Lagoze, Cornell University* Focus on plans for implementation, future research agendas, and community building.	

List of Participants

1. Antonia Arachova - National Library of Greece, Greece
2. Thomas Baron - Cern, Switzerland
3. Hans J. Becker - State And University Library Goettingen, Germany
4. Filipe Bento - University of Aveiro, Portugal
5. Giorgio Bertolla - Ires-Piemonte, Italy
6. Marvin Brunner - Océ-Technologies B.V., Netherlands
7. Gerhard Budin - University of Wien, Austria
8. Ines Cordeiro - Biblioteca De Arte-Fundação Calouste Gulbenkian, Portugal
9. Americo De Sousa Martins - Ineti / Dmg, Portugal
10. Timothy DiLauro - Johns Hopkins University, USA
11. Susanne Dobratz - Humboldt-University At Berlin, Germany
12. Paul Doorenbosch - Netherlands Institute For Scientific Information Services, Netherlands
13. Jorgen Eriksson - Danish National Library Authority, Denmark
14. Vanda Fidalgo – INESC, Portugal
15. Edward Fox - Virginia Tech, USA
16. Norbert Fuhr - University of Dortmund, Germany
17. Jorge Gustavo De Albuquerque Furtado Lopes - Entidade Reguladora do Sector Eléctrico, Portugal
18. Steve Griffin - US National Science Foundation
19. Rachel Heery – Ukoln, UK
20. Birgit Henriksen - The Royal Library, Denmark
21. Geneva Henry - Rice University, USA
22. Verlene Herrington - Ag Communication Systems (Lucent), USA
23. Eberhard R. Hilf - Institute For Science Networking, Germany
24. Linda L. Hill – University of California, Santa Barbara, USA
25. Sarantos Kapidakis - National Documentation Center, Greece (Hellas)
26. Laszlo Kovacs – MTA, Hungary
27. Thomas Krichel, RePEc, UK
28. Carl Lagoze - Cornell University, USA
29. Antoinette lemaire - Université catholique de Louvain, Belgium
30. Christopher Leonard - Elsevier Science, Netherlands
31. Kurt Maly - Old Dominion University, USA
32. Patricia Manson - EC - European Commission, Luxembourg
33. Cliff Morgan – John Wiley&Sons Ltd, UK
34. Eva Müller - Uppsala University Library, Sweden
35. Michio Obara - Japan Science And Technology Corporation (JST), Japan
36. Ann-Christin Persson - Lund Univ. Library, Netlab, Sweden
37. Thomas Place - Tilburg University, Netherlands
38. Isabel Ringel - CSIR Bio/Chemtek, South Africa
39. Jack A. Robinson - Dow Chemical, USA
40. Marta Rosete, Portugal
41. Diann Rusch-Feja - Max-Planck Institute For Human Development, Germany
42. Mogens Sandfaer - Technical Knowledge Center, Denmark
43. Rudi Schmiede - Darmstadt University of Technology/Global Info, Germany
44. Yoshiko Shirokizawa - Japan Science And Technology Corporation (JST), Japan
45. Ana Luisa Pereira Da Silva - Biblioteca Geral Univ. De Coimbra, Portugal
46. Bernard Smith - DGXIII, European Commission
47. Jadranka Stojanovski - Rudjer Boskovic Institute, Croacia
48. Shigeo Sugimoto - University of Library And Information Science, Japan (Nippon)
49. Robert Tansley – University of Southampton, UK
50. Ana Taveira - Universidade dos Açores, Portugal
51. Constantino Thanos - IEI-CNR, Italy
52. Yin Leng Theng - Middlesex University, UK
53. Harold Thimbleby - Middlesex University, UK
54. Eric Van De Velde - California Institute of Technology, USA
55. Stuart Weibel – OCLC, USA
56. Ingerborg Zimmermann - University of Zurich, Switzerland

Position Statements

Baron, Thomas

Goals and Issues at CERN:

Goal: How we view OAi:

- First goal is a more complete and costless harvesting of all HEP documents.
- All HEP institutes should be OAi-compliant (at least data providers).
- CERN will continue collecting and distributing the Open Archives of interest to its community.
- In parallel, CERN will become an OA for its public collection (e-prints but also photos, videos, internal notes, etc.). On demand, CERN could also be an OA for other collections (belonging to other institutes).
- In HEP, will we move from a central system to a distributed one (based at institutes)?

Waiting for the services:

- We can see a few collections declared as Open Archives but they do not really respond to the protocol (who will check this in the future?).
- We do not yet see any services at all. Why? What's the point of becoming OAi-compliant if no additional services are available?

Ownership of metadata and data?

- A service cannot pretend to own the data if the metadata is open.
- How is the origin of the data warranted when harvested in many places?
- What is the relation Author<-->his institute<-->the Open Archive?

Poor quality of metadata

- CERN librarians feel concern that the minimum level of metadata in OAi is too poor. The consequence of this is that the cost of processing (cleaning) the information will not be reduced (compared to XXX).
- Will there be a standard for field formatting?

Peer reviewing:

If all documents go through institutes, quality will be warranted (because the institute number will be the main reference). If this is not the case (as today), will peer reviewing enter into Open Archives Initiative's concerns?

Proposition: build Open Source software for Open Archives.

- All institutes could use it to become OAi-compliant.
- Is this already foreseen?

Fuhr, Norbert

Our general interest in Open Archives is the definition of standards for services and the implementation of these services.

Our group is involved in two projects related to Open Archives.

In the Carmen project, we are developing a new retrieval system for RDF metadata and XML documents. This system will be an extension of the Harvest system. Its primary application will be the preprint servers of the German societies of Mathematicians and Physicists. The system will collect document Metadata according to the Dublin Core standard, convert it into XML and then perform retrieval on XML documents. For the latter task, we are implementing a new retrieval engine with a query language that is an extension of XQL with information retrieval features. The system is Open Source, and a first prototype will be released to the public very soon.

The CYCLADES project will develop a service environment for open archives following a federative approach. In particular, the following services will be provided:

- Access service
- Browse service
- Collection service
- Personalisation service
- Recommendation service
- Collaborative Work service

For implementing the Access service, we will use the retrieval engine developed in the Carmen project.

Furtado Lopes, Jorge Gustavo

I have three points of discussion that I think are important for this meeting:

- The first is the question of the queries and capacities of the librarians of XXI century.
- The second is the safety and privacy of the personal information of the Public Digital Libraries.
- The last problem is the link between librarians and Internet development/Computer Professionals in building software solutions for the Library of this Digital Age.

Henry, Geneva

Executive Director for the Digital Library Initiative at Rice University.

Rice University's digital information environment is rich, complex and very heterogeneous. In an attempt to define a manageable and maintainable architecture for integrating the many useful electronic information resources, it is necessary to promote standards that will facilitate access to scholarly information by as many users as possible. Repositories of information will continue to remain distributed, with a variety of useful scholarly content added by users from all disciplines across the campus. This will include full text, video, audio, image and structured data authored and published with a variety of tools on a range of platforms.

I am interested in the efforts of the Open Archives Initiative to promote a common level of access to scholarly works, but I have many questions about how this should be promoted and implemented on the Rice campus. The XML standard is starting to be examined and adopted by more and more groups. It will be difficult to educate, train and expect local information providers to adopt yet another standard, after so many standards have seemingly come and gone over the past several years.

The improvement of search engines over the past few years also raises the question of how the enforcement of the Open Archives initiative will improve access to information when these newer commercial systems are demonstrating vastly improved performance for recall and precision across large repositories of a variety of diversely structured text and at fairly high speeds. The retrieval of information around concepts is getting to be quite common in the newer search engines, improving the quality of information retrieved for end users.

The following is a list of issues I will be seeking to evaluate in the discussions of this Open Archives initiative workshop:

1. What are the benefits of the Open Archives Initiative over other standards and search engines? How will OAi integrate with these standards and search engines?
2. How will information contributors be incentivized to conform to this standard if adopted? How widespread will the standard be?
3. How will harvesting engines (e.g. Dienst) be made available and supported over time?
4. How will this initiative support digital preservation? Also, how will digital preservation research be aided by this new standard?
5. If extended beyond the e-journal types of content, will the Open Archives initiative overcome barriers that other standards have been unsuccessful in overcoming? What will happen to these other standards?
6. Will there be limitations to the operating systems, protocols, etc. on which the harvesting engines will operate?
7. How well will this approach scale? Where does it break down? What is the proposed architecture to ensure scalability to accommodate large data sets and large repositories (into the pedabytes)?
8. What is the consequence of not adopting the Open Archives initiative across research universities?

Hilf, Eberhard R.

PhysDoc will join the OAI mission.

PhysDoc is, in the OAI terminology, both a data provider and a service provider. The database is a set of about 1000 distributed Physics Institutions with local documents and document information distributed around the world.

Metadata: an upload form is used (the same as on Math-Net), MMM (My Meta Maker), which creates the DC-metadata if the author fills them in.

The service is a distributed set of mirror-brokers, which search across the distributed databases using Harvest.

For further information: http://www.eps.org/PhysNet/

Statement for the development of OAi:

1. Allow for distributed local databases where the author stays responsible for the document.

2. Allow for services which search across these databases, without requiring that every little institute has to learn how to install DIENST or the like.

3. Discuss the metadata for OAi. We see them as very library-oriented and (possibly) too minimal.

We propose to start from the set used in PhysNet and Math-Net, which are DC-compliant, and have been tested and used in practice for some years.

4. A Metadata Upload form should be used and offered for all database servers, adapting the largely used MMM (My MetaMaker), making it independent of specific fields. The present version is 1.31 (see http://www.physik.uni-oldenburg.de/EPS/mmm/).

Sugimoto, Shigeo

University of Library and Information Science
Tsukuba, Japan

My current research interest is mainly in metadata, especially in subject gateways and metadata registries. Since the activities at the OAi include very important points for these topics, I'd like to participate in the workshop and share knowledge and information with other participants.

Firstly, I'd like to present some crucial issues from a general viewpoint of DL:

1. Medium/long term maintenance of metadata: Definitions of metadata elements and vocabularies will change over time. These changes should be properly maintained in a digital form in order to allow database/IR systems to be able to use them for interoperability between legacy and new data.
2. Interoperability among DLs: It is well known that interoperability among DLs is very important. Levels and types of interoperability should be clarified in order to enhance interoperability, eg., mirroring, cooperative contents/collection development, search protocols, metadata sharing, etc.
3. Metadata interoperability: Metadata is the crucial part of a DL because it is used for various purposes, for example, to find resources, to control access to the resources, to manage accounting, etc. Therefore, interoperability of metadata is the key aspect for interoperability of DLs. Metadata interoperability across languages is also important for international information resource access.
4. Inter-DL (Inter-Archive) collaboration for metadata development: Unified metadata databases of information resources will be very useful for information resource access on the Internet, e.g., union catalog of network information resources. In the current Internet environment, search engines and directory services, which use search robots to collect resources, are widely used on the Internet for finding information resources. However, on the other hand, quality resources stored in DLs tend to be stored in databases that are not accessible to the robots. This, in turn, means metadata sharing is crucial to enhance accessibility to the quality resources, and collaborative development of the shared metadata by DLs is important.

Secondly, I'd like to briefly present some viewpoints on subject gateways and metadata registries.

1. Subject Gateway (SG): An SG provides functions to navigate the users to appropriate information resources in a certain subject area(s). From the viewpoint of SGs, the academic information resources stored in the participating services of OAi are quite important resources. SGs would be able to provide navigation functions in their own ways, which are different from those given by the services. In a sense, SGs have potential to add value to the archives.
2. Metadata Registry: I've been involved in a Dublin Core metadata registry for multiple languages, which provides a set of reference descriptions of the Dublin Core Metadata Element Set written in multiple languages. A metadata registry is a service to store key information of a metadata schema, e.g., reference description of metadata description schema, guide-lines to write metadata entries, application profiles, etc. Metadata interoperability and long-term usability are crucial issues. To cope with these issues, metadata registries will play an important role.

Theng, Yin Leng and Harold Thimbleby

Our background is that we are computer scientists and usability people, and we are implementing a digital library for children. We are interested to participate in this workshop for the following reasons:

1. To share understanding of the design and usability issues (our project is a children's self-archiving digital library of stories and poems)

We are building a children's digital library (DL) of stories and poems by and for 11-14 year olds. If DLs are to be popular with children, they need to be fun, easy-to-use and empowering them — both as

readers and *authors*. Therefore, the children's digital library we are building aims to provide children with:

- the opportunity for creation of their own stories/poems and uploading them into the temporary DL space for reviews from their teachers and peers, before submitting to the permanent DL. Only stories and poems approved by the teachers can be submitted to the DL, thus ensuring the quality of the documents;
- the opportunity to encourage collaboration, children can read stories, give reviews, read other children's reviews on stories and email authors for other comments; and
- the possibility of their own temporary workspace and allowing their search items to be ordered and prioritised according to, for example, either the year of publication or by the collection or categories within the library space.

We are presently working closely with an English teacher, his class students, and a librarian in a secondary school in London to integrate the use of the children's digital library into the school curriculum. Issues arise in the implementation of the children's digital library, especially in the dissemination and growth of the children's digital library.

We would like to encourage more schools and individuals to join the children's digital library. In the future, we hope to extend its facilities to include (to name just a few):

- educational content and activities for students above 14 years;
- research and scholarly activities amongst academics and researchers; and
- professional writing activities amongst editors, journalists, etc.

2. To extend the understanding of the design and usability issues gained in our project to a wider context of the interoperability of federated digital libraries.

From our experience with the children's digital library and our background in HCI/usability, we hope to share and extend our understanding of the design and usability issues in the wider context of the interoperability of federated digital libraries.

3. To gain an understanding of what has been accomplished in the Open Archives Initiative.

We are new to the Open Archives Initiative and its technicalities. We would like to better understand the goals of the Open Archives Initiative to examine whether and to what extent they are relevant to our need.

Thimbleby, Harold

(see Theng, Yin Leng)

Van de Velde, Eric

At Caltech, we are currently well under way in three distinct Open Archives projects:

1. Electronic Theses and Dissertations

We are a member of NDLTD, and intend to make electronic submission of all theses compulsory in the near future. At this time, we have installed the NDLTD software. We are working with the graduate dean's office on submission of theses.

2. Departmental Technical-Report Archives

For any Caltech department that wants it, the Caltech Library System offers the service of maintaining a repository of technical reports. Each department individually decides on major policies regarding submission standards and procedures.

Since most faculty have not thought about such policies in great detail, the library helps them formulate policies for their departmental repositories. The library imposes very few policies of its own. Thus far, four library-imposed policies have emerged as a result of these discussions:

- Submission of the documents into the repository must not violate applicable copyright restrictions.
- To ensure stability and permanence of the archive, submitted documents cannot be withdrawn, but they can be updated.
- A librarian who makes a final quality control check on format and metadata does the final step of submitting a document into the archive.
- The library will not track who reads/downloads which documents.

As our recruitment broadens, we may be faced with other requests that we cannot honor. However, we intend to keep the number of library-imposed policies to an absolute minimum.

We have currently one department (Computer Science) fully active and integrated into the NCSTRL federation. We are working with other departments on developing submission standards. Internally, we are training all librarians in the document-management skills necessary to perform these new duties.

Eventually, we want all of these repositories to be federated through the Open Archives initiative into disciplinary federations (like NCSTRL). However, we are also interested in running our own federations: we want to federate departmental repositories into divisional federations and into one Caltech-wide federation.

3. Electronic Journals

We have received several requests from Caltech faculty members for help with putting up electronic journals and/or conference proceedings. We believe this is the start of a big trend. Faculty interested in new journals will not even approach commercial enterprises, preferring to maintain control over the whole process. The library intends to provide editorial-management services to these fledgling editorial boards. Obviously, we want the editorial management to be as automated as possible. In addition, we want to make available these journals and proceedings through the Open Archives initiative.

In surveying the field, we found a lot of excellent software, and we decided that we would not develop our own software. Instead, we decided to focus on the organizational aspects and use available free software to the maximum extent possible. However, we are encountering some obstacles with this approach, and I believe that it is important to bring them up for discussion at the Open Archives meeting.

The first issue is that we should not underestimate the cost and time required to educate our user communities. Students, faculty, librarians, and support staff must be trained to use the document-management systems that perform:

- Editorial workflow, including initial submission, review by editor(s), referee(s), copy editor(s), technical-format editor(s), and metadata editor(s)
- Submission to the archive
- Federation through Open Archives
- Provide the ability to customize the interface to the local repository and to federated archives that combine selected local repositories.

It is crucial that these enabling technologies be user friendly, focused not on programmers but on the end users. This will ease the recruitment of departments and faculty into this scholarly-communication initiative.

The second issue revolves around the ability to run federations. Software like the NCSTRL collection service, which allows one to federate many individual repositories, needs to become widely available. Discipline-oriented repositories like the LANL arXiv and federations like NCSTRL were crucial in freeing the scholarly literature in specific areas. However, if our initial work at Caltech is any indication, this approach may not be scalable. Most departments and individual faculty want to retain a certain level

of control over these repositories. That is why we are developing different repositories with policies tailored for different administrative units of Caltech. For example, there will be at least two different Computer Science repositories: one for the Computer-Science department and one for the Center for Advanced Computing Research (an on-campus independent research institute). One single repository per discipline is not feasible, even within a small campus like Caltech. Instead, the repository boundaries are far more arbitrarily defined by the vagaries of administrative structures: one repository for each authority.

As we broaden our work to encompass more disciplines, we may want to run our own NCSTRL-like federations. For example, suppose a departmental repository in discipline X becomes popular. It is not inconceivable that colleagues on other campuses want to join the effort. In that case, we would like to offer running international discipline-X-oriented federations (like NCSTRL) at Caltech.

Weibel, Stuart

My interests in the OAi Workshop are motivated by two concerns:

1. Metadata

As director of the Dublin Core Metadata Initiative, I am eager to see close cooperation between DCMI and the OAi, to promote as best we can, a coherent metadata position that will meet the functional requirements of OAi and be interoperable as far as possible with existing and evolving DC recommendations.

2. Library infrastructure

As a representative of the OCLC Office of Research, I see OAi as an important part of the emerging infrastructure for the dissemination and maintenance of the scholarly publishing record, and we would be pleased to see a collaborative research activity develop if this meets mutual needs.

D-Lib Magazine
February 2000

Volume 6 Number 2
ISSN 1082-9873

The Santa Fe Convention of the Open Archives Initiative

Herbert Van de Sompel
Los Alamos National Laboratory - Research Library, New Mexico, US, and
Automation Department of the Central Library of the University of Ghent, Belgium
herbert.vandesompel@rug.ac.be.

Carl Lagoze
Department of Computer Science
Cornell University
lagoze@cs.cornell.edu.

Abstract

The Open Archives initiative (OAi) promotes and encourages the development of author self-archiving solutions (also commonly called e-print systems) through the development of technical mechanisms and organizational structures to support interoperability of e-print archives. Such interoperability can stimulate the transition of e-print systems into genuine building blocks of a transformed scholarly communication model. This paper describes the Santa Fe Convention of the OAi. This is a set of relatively simple but potentially quite powerful interoperability agreements that facilitate the creation of mediator services. These services combine and process information from individual archives and offer increased functionality to support discovery, presentation and analysis of data originating from compliant archives.

Introduction

In July 1999, Paul Ginsparg, Rick Luce and Herbert Van de Sompel sent out a Call for Participation (Ginsparg, Luce, and Van de Sompel 1999a) to a meeting exploring cooperation among scholarly e-print archives. The meeting, held in October 1999 in Santa Fe, and originally called the Universal Preprint Service meeting, led to the establishment of the Open Archives initiative (OAi) (Ginsparg, Luce, and Van de Sompel 1999b). The goal of the OAi is to contribute in a concrete manner to the transformation of scholarly communication. The proposed vehicle for this transformation is the definition of technical and supporting organizational aspects of an open scholarly publication framework on which both free and

commercial layers can be established.

This paper describes the origins of the OAi and work heretofore in defining this framework: the Santa Fe Convention. This convention is a combination of organizational principles and technical specifications to facilitate a minimal but potentially highly functional level of interoperability among scholarly e-print archives. The convention gives *data providers* -- individual archives -- relatively easy-to-implement mechanisms for making information in their archives externally available. This external availability then makes it possible for *service providers* to build higher levels of functionality, *mediator services*, using the information made available from scholarly archives that adopt the convention.

The growth of e-print archives

The origins of the Open Archives initiative lie in the growing number of electronic preprint (e-print) archives. While several of these began as informal vehicles for the dissemination of preliminary results and non-peer reviewed "gray literature", a number of them have evolved into an essential medium for sharing research results among the colleagues in a field.

These archives demonstrate a shift in the traditional scholarly communication model, which has relied on formally published scholarly journals. There is a growing consensus that the scholarly journal system is facing significant challenges:

- The explosive growth of the Internet has given scholars almost universal access to a communication medium that facilitates immediate sharing of results.
- The rapidity of advances in most scholarly fields has made the slow turn-around of the traditional publishing model an impediment to collegial sharing.
- The full transfer of rights from author to publisher often acts as an impediment to the scholarly author whose main concern is the widest dissemination of results.
- The current implementation of peer-review -- an essential feature of scholarly communication -- is too rigid and sometimes acts to suppress new ideas, favor articles from prestigious institutions, and cause undue publication delays.
- The imbalance between skyrocketing subscription prices and shrinking or, at best, stable library budgets is creating an economic crisis for research libraries.

The e-print archives exemplify a more equitable and efficient model for disseminating research results. An important challenge is to increase the impact of the e-print archives by layering on top of them services -- such as peer review -- deemed essential to scholarly communication. This is the focus of the Open Archives initiative.

An exhaustive review of existing e-print archives is out of the scope of this paper. An interesting list of initiatives is available at the Office of Scientific and Technical Information. A brief review of some of the notable efforts is illustrative of the scope of these initiatives:

- arXiv.org, hosted by Los Alamos National Laboratory, is considered the premier example of e-print archives. The archive was started in 1991 by Paul Ginsparg, who is internationally recognized as one of the leaders in the area of scholarly publishing alternatives. Over the past decade, the arXiv archive has evolved towards a global repository for non peer-reviewed research papers in a variety of physics research areas.

arXiv has also incorporated mathematics, non-linear sciences and computer science.

- CogPrints, hosted by the University of Southampton in the U.K., is modeled on arXiv and focuses mainly on papers in Psychology, Linguistics and Neuroscience.
- NCSTRL (Networked Computer Science Technical Reference Library) is an international collection of computer science research reports. NCSTRL is based on a distributed model. Documents are stored in distributed archives and are made available through distributed services that communicate via the Dienst protocol.
- NDLTD aims at building a digital library of electronic theses and dissertations (ETD) authored by students of member institutions. In ongoing research, NDLTD addresses issues such as the creation of a workflow to submit ETDs, the development of an XML DTD for ETDs and the support of a digital library for ETDs.
- RePEc, an initiative in economics, also operates on a distributed model. It provides authors with the option to submit working papers to a departmental archive or -- if one does not exist -- to the EconWPA archive at Washington University. These archives support the so-called Guildford protocol that guarantees interoperability between the RePEc archives and has enabled the creation of a variety of end-user services.

There are indications that a growing number of disciplines and organizations are inspired by this pioneering work and are investigating alternative models for scholarly communication:

- The NIH e-biomed proposal (Varmus 1999) for a more effective communication system for research reports in the life sciences demonstrates the innovative thinking inspired by initiatives like arXiv. While the PubMed Central environment (Anonymous 1999) (the system being developed as the outcome of the proposal) is more conservative than e-biomed, it remains faithful to the original desire to provide barrier-free access to primary reports in the life sciences.
- The British Medical Journal and HighWire Press recently launched Clinical Medicine Netprints (Delhamothe, et al. 1999), an e-print site for studies, research, and articles in Clinical Medicine.
- Under the umbrella of the eScholarship project, the California Digital Library is working on University ePub (Lucier and Ober 1999), a set of disciplinary e-print servers and services whose overall aim is to lead and support innovations in the production and dissemination of scholarship. The project received one of the grants from SPARC in the context of its Scientific Communities Initiative, which called for proposals introducing alternative communication methods as a way to address the serials crisis.
- MIT plans to build a digital repository of which all public e-prints will be available to the whole e-print community.
- Caltech's Scholar's Forum (Buck, Flagan, and Coles 1999) describes an alternative conceptual model for scholarly communication.

From individual archives to an interoperable fabric

The aim of the archive initiatives described above is to try to create a more effective scholarly communication mechanism that addresses problems that exist in the established system. The approaches that are taken by individual archives differ in a number of ways. Some initiatives

build on a centralized model, others on a distributed departmental, or by extension, institutional model. Some deal with gray (non-peer reviewed) literature only, others incorporate metadata of peer-reviewed papers or try to establish some form of peer-review outside of the established system. Some deal with metadata only, others with both metadata and full content. Yet all share the attribute of offering scholars a vehicle to conveniently and immediately disseminate research results to peers.

The reason for launching the Open Archives initiative is the belief that interoperability among archives is key to increasing their impact and establishing them as viable alternatives to the existing scholarly communication model. This conviction is expressed in the official mission statement of the initiative:

The Open Archives initiative has been set up to create a forum to discuss and solve matters of interoperability between author self-archiving solutions (also commonly referred to as e-print systems), as a way to promote their global acceptance.

Interoperability is a broad term, touching many diverse aspects of archive initiatives, including their metadata formats, their underlying architecture, their openness to the creation of third-party digital library services, their integration with the established mechanism of scholarly communication, their usability in a cross-disciplinary context, their ability to contribute to a collective metrics system for usage and citation, etc.

Interoperability among archives offers substantial benefits to the scholars that use them. An important attribute of the traditional research library as an information provider is its role as a common entry point for a variety of information resources, not necessarily divided along disciplinary or institutional boundaries. The move from physical to digital sources should not be accompanied by the breakup of this entry point into a collection of fragmented archives. An increasing number of scholars move fluidly in their research across domain boundaries; the technology for delivering digital information should facilitate rather than hinder such fluidity. Mechanisms for interoperability offer the potential for discovery tools and virtual collections (Lagoze, 1998) that extend across the contents of multiple archives. Authors also benefit from such archive spanning tools, since their works will be accessible by a wider audience.

Interoperability is also beneficial to the archive and service provider. Rather than having to provide an entire suite of services for its users, individual archives can instead establish a well-defined interface on which external providers can build enhanced services. A variety of such services can be envisioned, including those that facilitate discovery, linking, and reviewing. An intriguing and essential set of services would be those that provide metrics to assist in the evaluation of the impact of certain scholarship and aid in tenure review and promotion decisions.

The Sante Fe Convention of the OAi represents a pragmatic, incremental, and collaborative approach towards interoperability. The initiators of the Open Archive initiative hope that this practical approach will be a catalyst for significant changes in the mechanisms for scholarly communication. The need for such change has been the issue of numerous papers, workshops, and Internet discussion groups. Yet, the existing system has proven somewhat resistant to change, no doubt due to the complex socio-political and economic forces that support it. For example, the current system of academic promotion and tenure is closely linked to the traditional journal system (Wilson 1942). This acts as an important factor sustaining the existing communication model (Schauder 1994). Understandably, scholars are hesitant to support alternative models that are not yet linked to their evaluation and promotion. While such issues will continue to support the current system, the development of practical technical and organizational solutions, such as the Sante Fe Convention, builds a framework for changes that

will inevitably occur and may encourage the implementation of those changes.

Agreeing on interoperability: the Santa Fe meeting of the Open Archives initiative

A successful first meeting of the initiative was held on October 21-22, 1999, in Santa Fe, New Mexico. The meeting was sponsored by the Council on Library and Information Resources (CLIR), the Digital Library Federation (DLF), the Scholarly Publishing & Academic Resources Coalition (SPARC), the Association of Research Libraries (ARL) and the Los Alamos National Laboratory (LANL). The participants were computer scientists and digital librarians. There were also representatives of existing and emerging e-print systems, of scholarly publishers and of the sponsors. All but one of the invited institutions sent a representative. This was considered to be a firm indication of the perceived importance of the initiative.

The central theme of the first meeting was the establishment of recommendations and mechanisms to facilitate cross-archive value-added services. Such services could combine information derived from cooperating archives, process that information to produce some value-added information, and make that enhanced information available to users, agents, or other services. Examples of such services include cross-archive search engines, current awareness services, linking systems, and peer-review services.

Achieving progress on this goal required agreement among the participants on the issue of interoperability. Although interoperability has been a watchword for a variety of efforts in digital libraries and networked information (Paepcke, Chang, et al. 1998), the actual meaning of it and the implementation thereof has often proven elusive. Like many meetings intended to reach agreement on standards, attendees at the Santa Fe meeting arrived with a variety of pre-conceived notions on what was required to reach interoperability. It is instructive to review how these differing notions converged into a well-defined agreement that provides the foundation for cross-archive exchange of information.

The meeting began with a rather expansive example of interoperability, illustrated through the UPS Prototype project coordinated by Herbert Van de Sompel, Thomas Krichel, and Michael Nelson. This project and its results are described at length in the companion paper (Van de Sompel, Krichel, Nelson, et al 2000). Briefly summarized, the prototype demonstrated the integrated operation of a variety of services operating over data originating from a set of archives. Each of those services provided a reasonably rich level of functionality (implemented through a set of protocols).

There was general agreement among the participants at the meeting that the Prototype was an extremely useful demonstration of potential. There was also agreement, however, that trying to reach consensus on the full functionality of the Prototype was "aiming too high" and that a more modest first step was in order. The Prototype team, based on their insights gained during implementation of the UPS prototype, also reached a similar conclusion. This is described more fully in "Recommendations made to the Open Archive group" of (Van de Sompel, Krichel, Nelson, et al. 2000).

The remainder of the meeting was engaged in determining the proper degree of modesty, which balanced the need for adequate functionality against the requirement that the cost of entry for participating archives be sufficiently low. This is a question that has bedeviled other efforts at interoperability; for example, buy-in to the highly functional Z39.50 protocol has largely been limited to libraries, due to the costs of complexity (Stubley 1999). An important step towards establishing the cost/functionality balance was reached by the beginning of the second day with agreement among the participants on a tiered model of interoperability. This model is illustrated

in Figure 1, showing the following layers:

- *Document Models* - that address document structure and allow the specification of multiple disseminations (e.g., in multiple formats or of various structural decompositions) of a document instance. One example that addresses this level of interoperability is the Dienst repository protocol.

- *Metadata Harvesting* - that enables the extraction of descriptive surrogates for documents. This approach was effectively demonstrated by the Harvest project (Bowman 1995) several years ago.

- *Mediator Services* - that describe the nature of services that use and enhance information available from archives. The UPS Prototype (Van de Sompel, Krichel, Nelson, et al 2000) demonstrated a number of these services (e.g., linking) that build on top of the metadata harvesting layer. This service layer is also described in the digital library service model of (Leiner 1998).

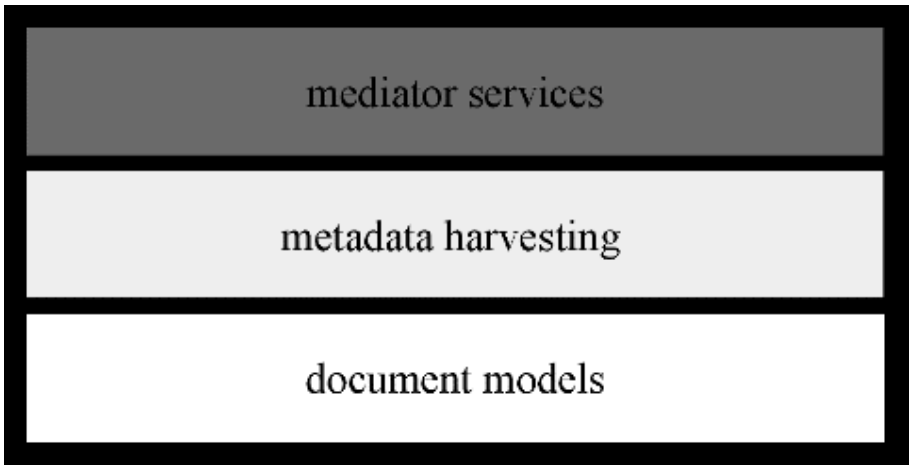


Figure 1: a tiered model of interoperability

Framing the problem of interoperability with this model quickly led to the decision to restrict the Santa Fe recommendations to interoperability at the level of *metadata harvesting*. The mechanisms for establishing this interoperability, described in full detail in the Santa Fe Convention and summarized in the remainder of this paper, are three-fold:

1. The definition of a set of simple metadata elements -- the Open Archives Metadata Set (OAMS) -- for the sole purpose of enabling coarse granularity document discovery among archives;

2. The agreement to use a common syntax, XML, for representing and transporting both the OAMS and archive-specific metadata sets;

3. The definition of a common protocol -- the Open Archives Dienst Subset -- to enable extraction of OAMS and archive-specific metadata from participating archives.

This agreement treats documents as black-boxes; archives can have idiosyncratic document representations with the Santa Fe Convention only specifying a URL entry point to the archives' individual document models. The question and functionality of common mediator services are left open to implementers who wish to exploit the Santa Fe Convention and build mechanisms

based on it.

The Santa Fe Convention

Objective

The Santa Fe Convention presents a technical and organizational framework designed to facilitate the discovery of content stored in distributed e-print archives. It makes easy-to-implement technical recommendations for archives that –- when implemented –- will allow data from e-print archives to become widely available via its inclusion in a variety of end-user services such as search engines, recommendation services and systems for interlinking documents. In addition, the convention introduces an organizational framework for making information available about archives that adhere to the technical recommendations of the convention -- the *data providers* -- and about trusted parties that build end-user services for data originating from such archives -- the *service providers*. As such it provides a communication mechanism between providers of data and providers of services and creates a community of open archives.

Definitions and Concepts

The Santa Fe Convention builds on on a number of definitions and concepts that are essential for its understanding.

Open and managed e-print archives

The Convention considers the following to be crucial components of an e-print archive:

- A submission mechanism;
- A long-term storage system;
- A management policy with regard to submission of documents and their preservation;
- An open machine interface, that enables third parties to collect data from the archive.

The last item is crucial for enabling third parties to create services that support the discovery, presentation and analysis of data in the archive. Most e-print archives will also provide native end-user services. However, facilitating the broad dissemination of archive data through third party services is a crucial feature of an e-print archive. Therefore, the open interface is a key part of the Santa Fe Convention.

Data providers and service providers

Consistent with the objective of the Santa Fe Convention and the identification of the crucial functions of an e-print archive, there is a distinction between two participants in the convention:

- A *data provider* is the manager of an e-print archive, acting on behalf of the authors submitting documents to the archive. As pointed out above, the data provider of an open archive will, at least, provide a submission mechanism, a long-term storage system and a mechanism that enables third parties to collect data from the archive;
- A *service provider* is a third party, creating end-user services based on data stored in e-print archives. For instance, a service provider could implement a search engine for mathematical e-prints stored in archives worldwide.

Data in an e-print archive

The convention uses the notion of a *record* in an archive. Some archives may store metadata that describes full content without storing the full content itself. In this case, the metadata is a *record*. Other archives may also store full content. However, the convention assumes that if full content is stored, there will always be associated metadata stored in the archive as well as a mechanism to tie metadata and content together. In this case the combination of metadata and full content is a *record*.

Technical Components of the Santa Fe Convention

The complete details of the technical components of the Santa Fe Convention and instructions for participating are available via the core document. Organizations considering participation should refer to that document. This section summarizes the information for the purpose of an overview.

Open Archives Metadata Set

The Open Archives Metadata Set (OAMS) is a collection of nine metadata elements intended to facilitate coarse granularity resource discovery among the records in distributed and dissimilar archives. The semantics of this set have purposely been kept simple in the interest of easy creation and widest applicability. There is no provision for qualification or extension of the nine elements. The expectation is that individual archives will maintain metadata with more expressive semantics and the Open Archives Dienst Subset provides the mechanism for retrieval of this richer metadata.

Open Archives Dienst Subset

The Open Archives Dienst Subset is a set of protocol requests that are delivered via HTTP. This protocol is a subset of the full Dienst protocol. The protocol requests in the subset provide the following functionality:

- List the full identifiers for records stored in an archive. An optional argument permits the client to specify that the list should only include records added after a specific date. Another optional argument allows the client to specify that the records should be accompanied by the metadata associated with the identifier.
- Return the metadata for a specific record in a requested format.
- Return the list of metadata formats supported by an archive.
- Return the list of metadata formats available for a specific record.
- Return the structure of the partitions by which an archive is organized.

All responses to these requests are formatted in XML.

Organizational aspects of the Santa Fe Convention

The convention also introduces an organizational framework to facilitate its implementation and to establish a communication mechanism between data providers and service providers. An understanding of this framework can be obtained from an exploration of the core document of the Santa Fe Convention that gives a step by step approach for making an e-print archive or a service comply with the Santa Fe Convention.

For the data providers, some of these steps are directly related to the implementation of the technical recommendations of the convention, as summarized in the previous section. In addition, the core document introduces the following important organizational elements:

- The notion of a *unique archive identifier* to unambiguously represent each e-print archive, as well as the notion of a *full identifier* of a record in an archive. Since this *full identifier* is a concatenation of the unique archive identifier and the unique persistent identifier of a record in an archive, this *full identifier* will be persistent and globally unique.
- The recommendation to document metadata formats other than the OAMS, as well as a facility to share this documentation with data providers and service providers.
- A facility to register an e-print archive as being compliant with the Santa Fe Convention, by means of the provision of a filled-out version of a data provider template that describes crucial characteristics of the archive. Important information to be provided in this template is the *unique archive identifier*, the metadata formats implemented by the archive and URLs of the Dienst interface of the archive. The template also provides a means to provide information on the content of an archive, its submission policy and contact addresses. In addition, the template gives data providers a means to express the terms and conditions of usage of archive data.
- A list of e-print archives that comply with the Santa Fe Convention, from which links are available to the documents describing their crucial characteristics.

For the service providers, the steps described in the core document of the Santa Fe Convention introduce the following:

- The request to maintain the original *full identifiers* of the records harvested by the service provider.
- The request to comply with the terms and conditions that data providers have brought forward in their filled-out data provider template.
- A facility to register a service as being compliant with the Santa Fe Convention, by means of the provision of a filled-out version of a service provider template that describes aspects of the service. Amongst others, the service provider must mention from which archives information is being harvested as well as the fact that harvesting is compliant with the terms and conditions expressed by the data providers.
- A list of services that comply with the Santa Fe Convention, from which links are available to the filled-out templates that describe them.

Conclusions and future plans

The technical results of the Santa Fe meeting may be perceived as quite modest, and indeed they are. However, the technical moderation should be viewed in a broader context. First, it played an important role in bringing the Santa Fe meeting to a successful conclusion, with agreement among diverse parties. This agreement amongst a core group is an important step towards the development of a broader e-print community with a strong focus on cooperation and interoperability. The organizational framework provided by the Santa Fe Convention is intended to actively contribute to the creation and extension of such a community. Second, the limited

nature of the technological requirements lowers the cost of entry for new participants, and hopefully builds momentum for the development of scholarly publishing alternatives. This momentum will provide a basis for future agreements that may extend and enhance the current Santa Fe Convention.

If successful, the Convention will attract early adoption by existing archives and encourage the establishment of new scholarly archives that will support the mechanisms defined by the Convention. The former, early adoption, seems to be occurring with participants at the meeting representing arXiv, the California Digital Library, clinmed, CogPrints, RePEc and NCSTRL, stating their intention to comply with the Santa Fe Convention in the near future.The CogPrints team at Southampton also work on the implementation of a free software for e-print archives that will comply with the Santa Fe Convention (Harnad 1999). Based on the number of inquiries received since the Santa Fe meeting, there are reasons to be optimistic regarding the establishment and adoption by other existing and planned archives. Positive feedback has been received from representatives of German mathematical and physical e-print archives. In addition, several commercial and non-commercial parties have expressed interest in creating mediator services once archives have implemented the convention.

The current challenge for the Open Archive initiative is to maintain a focus on the successful dissemination and implementation of the Santa Fe Convention. Before considering whether it is necessary or appropriate to expand the nature of the interoperability agreements, it is essential that the mechanisms described in the current convention be widely implemented and tested in practice. Without such proof of concept, the initiative may find itself increasing the complexity (and cost of implementation) of the interoperability mechanisms without discovering if, in fact, the level of interoperability defined by the existing Santa Fe Convention is sufficient and practical. Any future work to expand the scope of the OAi should understand that the success of any interoperability standard must be measured relative to both its functionality and its cost of adoption (Arms 2000).

The near-term plans for the Open Archive initiative include public dissemination of the Santa Fe Convention scheduled for February 15, 2000, and meetings to review progress and chart future activities. This paper represents the initial public dissemination and the Open Archives web site will serve as a persistent and official record of the convention. The next meeting will take place at ACM Digital Libraries 2000 in San Antonio, Texas, in June 2000. The exact dates and place of this meeting will be posted on the Open Archives web site nearer to the June date. A European meeting is tentatively planned in conjunction with ECDL 2000 in Lisbon, Portugal, in September 2000.

References

Anonymous. PubMed Central: An NIH-Operated Site for Electronic Distribution of Life Sciences Research Reports. August 1999. [http://www.nih.gov/welcome/director/pubmedcentral/pubmedcentral.htm].

Arms, William Y. 2000. Digital Libraries. MIT Press.

Bowman, C.M., et al. 1995. The Harvest Information Discovery and Access System. Computer Networks and ISDN Systems. 28 no. 1 & 2: pp. 119-125.

Buck, Anne M., Richard C. Flagan, and Betsy Coles. Scholars' Forum: A New Model For Scholarly Communication. March 1999. [http://library.caltech.edu/publications/scholarsforum/].

Delhamothe, Tony and others. 1999. Netprints: the next phase in the evolution of biomedical publishing. British Medical Journal 319: 1515-6. [http://www.bmj.com/cgi/content/full/319/7224/1515].

Ginsparg, Paul, Rick Luce, and Herbert Van de Sompel. Call for participation in the UPS initiative aimed at the further promotion of author self-archived solutions. July 1999a. [http://www.openarchives.org/ups-invitation-ori.htm].

Ginsparg, Paul, Rick Luce, and Herbert Van de Sompel. the Open Archives initiative. July 1999b. [http://www.openarchives.org/].

Harnad, Stevan. 1999. Free at Last: The Future of Peer-Reviewed Journals. D-Lib Magazine 5, no. 12. [http://www.dlib.org/dlib/december99/12harnad.html]

Judson, Horace Freeland. 1994. Structural Transformations of the sciences and the end of peer review. Journal of the American Medical Association 272, no. 2: 92-4.

Lagoze, Carl. 1999. Defining Collections in Distributed Digital Libraries, D-Lib Magazine 5, no. 11. [http://www.dlib.org/dlib/november98/lagoze/11lagoze.html]

Leiner, B.M. 1998. The NCSTRL Approach to Open Architecture for the Confederated Digital Library. D-Lib Magazine 5, no. 12. [http://www.dlib.org/dlib/december98/leiner/12leiner.html]

Lucier, Richard and John Ober. Scholar-led Innovation in Scholarly Communication University ePub: An initiative in Electronic Scholarship. October 1999. [http://www.cdlib.org/eschol/summary.html].

Paepcke, Andreas, Chen-Chuan Chang, Hector Garcia-Molina, and Terry Winograd. 1998. Interoperability for Digital Libraries Worldwide. Communications of the ACM 41, no 4.

Schauder, Don. 1994. Electronic publishing of professional articles: attitudes of academics and implications for the scholarly communication industry. Journal of the American Society for Information Science 45, no. 2: 73-100.

Stubley, Peter. 1999. Clumps as Catalogues. Ariadne no. 22. [http://www.ariadne.ac.uk/issue22/distributed/distukcat2.html].

Van de Sompel, Herbert, Thomas Krichel, Michael L. Nelson and others. 2000. The UPS Prototype: An Experimental End-User Service across E-Print Archives. D-Lib Magazine 6, no. 2.[http://www.dlib.org/dlib/february00/vandesompel-ups/02vandesompel-ups.html].

Varmus, Harold. E-BIOMED: A Proposal for Electronic Publications in the Biomedical Sciences. May 1999. [http://www.nih.gov/welcome/director/pubmedcentral/ebiomedarch.htm].

Wilson, L. 1942. The academic man: a study in the sociology of a profession. New York: Oxford University Press.

Acknowledgements

The authors wish to thank:

- All the participants at the Open Archives meeting in Santa Fe in October, 1999. The hard work of all of these people made the results described here possible.
- Clifford Lynch and Don Waters for effectively chairing the Santa Fe meeting.
- Caroline Arms, Mark Doyle, Ed Fox, Paul Ginsparg, Thomas Krichel and Michael Nelson for their contributions to the Santa Fe Convention.
- CLIR, SPARC, ARL, and the LANL Research Library for financial and moral support without which the Santa Fe meeting would not have been possible.
- Donna Berg of the LANL Research Library for the perfect organization of the meeting.

Herbert Van de Sompel wishes to thank the Belgian Science Foundation for a special Ph.D. grant.

Work on Dienst and the Open Archives Dienst Subset is supported by the National Science Foundation Grant No. IIS-9817416 and Defense Advanced Projects Agency Grant No. N66001-98-1-8908, with the Corporation for National Research Initiatives.

(The link for clinmed was corrected to <http://clinmed.netprints.org> on 2/22/00 at the author's request.)

Report on Open Archives Initiative Technical Committee Meeting

Ithaca NY, 7-8 September 2000

Context

The original aim of the Open Archives Initiative was to provide an infrastructure for interoperability among sites supporting author self-archiving and thereby promote their wide acceptance. Although the Initiative generally concentrated on technical matters, its mission reflected its roots in the e-print community and the underlying political agenda to promote the ongoing transformation of scholarly communication. The inaugural meeting of the Open Archives Initiative (OAI) in October 1999 spawned an agreement now known as the Santa Fe Convention.

The Santa Fe Convention is a set of relatively simple interoperability agreements that facilitate a minimal but potentially highly functional level of interoperability among scholarly e-print archives through metadata harvesting. The interoperability agreements are a combination of organizational principles and technical specifications The Convention gives *data providers* -- individual archives -- relatively easy-to-implement mechanisms for making metadata in their archives externally available. This external availability then makes it possible for *service providers* to build higher levels of functionality, mediator services, using the information made available from scholarly archives that adopt the convention. These services may combine and process information from individual archives and then may offer increased functionality to support discovery, presentation and analysis of data originating from compliant archives.

Since the publication of the Santa Fe Convention in February 2000, interest has emerged from other communities who are interested in applying the framework for a wide variety of scholarly materials beyond e-prints. In order to respond to this wider interest, the OAI undertook a number of actions:

- The technical specifications were reconsidered in response to comments that certain aspects were e-print specific. Experimentation and discussion in the original e-print community also identified elements in the original specifications that required reconsideration.
- The original e-print specific mission statement was reconsidered. Rather than focusing on a political agenda focusing on author self-archiving, OAI's mission was reformulated to supply and promote an application independent technical framework - a supportive infrastructure that empowers different scholarly communities to pursue their own interests in interoperability in the technical, legal, business, and organizational contexts that are appropriate to them.
- Organizational changes were instituted to provide stability and credibility to the wider community base. A Steering Committee was appointed with the task of overseeing the pursuit of the mission. The activities of the Steering Committee will receive support from both the Digital Library Federation and the Coalition for Networked Information. In addition a Technical Committee was formed to focus on generalization and stabilization of the technical framework.

The Cornell Meeting of the OAI Technical Committee

A meeting of the OAI Technical Committee was held on September 7-8 2000 at Cornell University to revise the Santa Fe Convention in light of the changed context. The meeting included analysis of the current and emerging use of the interoperability framework and initiated the process of upgrading it to better serve the needs of a more general user base.

The meeting set out by agreeing on the following issues:

- The OAI interoperability framework should no longer only be concerned with e-prints, but with scholarly data-archives in general.
- Most fundamental principles of the Santa Fe Convention [open, harvestable archives ; data provider & service provider model ; managed archives] can be maintained in the extended scope. One concept [the definition of a record in an archive] should be reconsidered during the meeting.
- Most abstract principles that are presented in the Santa Fe Convention [metadata harvesting ; OAI namespace ; acceptable use ; registration of data providers, service providers and metadata formats] can

be maintained in the extended scope. One concept [shared metadata set & parallel metadata sets] should be reconsidered during the meeting.

- The existing technical implementation of these abstract principles should be reconsidered during the meeting because of the extension of scope and because of experiences with actual implementations.

The goal of the meeting was development of a new set of technical guidelines for consideration by the Open Archives Steering Committee and ultimate public dissemination by the beginning of 2001. Recognizing that any specification is subject to review and refinement, the attendees attempted to develop specifications that were:

- Stable for experimentation;
- Low risk for early adopters;
- Sufficiently easy implement so as to optimize the chances for future interoperability across communities.

It was decided to discontinue the name "Santa Fe Convention". The new name for the interoperability specification is "the Open Archives Harvesting Framework Specifications".

A formal specification document is currently being developed. The new specifications will be disseminated to the public in January 2001. Documentation, accompanying tools and software will be produced in parallel.

The remainder of this document summarizes the Open Archives Harvesting Framework Specifications. It focuses on issues where the Open Archives Harvesting Framework differs from the specifications in the Santa Fe Convention.

Record in an archive

The ambiguity in the original agreement about the definition of a *record* has been clarified. A record in an archive has been defined to be a metadata-record. The metadata record describes – and can contain an entry point to – full-content.

Metadata

The requirement to have a shared, basic metadata set to facilitate interoperability across communities was reconfirmed. Also the notion of parallel metadata sets that serve specific needs of communities and archives was reconfirmed.

The shared metadata set developed during the original Santa Fe meeting -- the OAMS -- was deemed inappropriate for cross-community use due to some e-print specific aspects. Instead, the Dublin Core Element Set was selected as the common metadata set. This selection leverages years of work in the Dublin Core Metadata Initiative in developing cross-community consensus.

Initial steps were taken to encourage the development of community-specific harvestable metadata sets. Representatives of the e-print community at the meeting decided to propose a metadata set targeted at the e-print community under the name EPMS by the beginning of 2001. Representatives of the research library community proposed a similar effort and calls for proposals from other communities (e.g., the museum community, Open Language Archives) will be issued.

To distinguish between metadata specific to harvesting functionality and other metadata (both shared and community specific), a carrier syntax in XML was developed. This syntax transports packages of specific sets (e.g., Dublin Core, EPMS) within a contextual wrapper that contains metadata specific to the harvesting interactions between data and service providers.

Identifiers and an OAI namespace

The concept of unique identifiers within an OAI namespace has been maintained. Its implementation has been revised for compliance with general URI principles and to allow for the building of resolution mechanisms and services across OAI-compliant archives. The following identifier syntax is proposed:

full-identifier = oai : archive-identifier : record-identifier

Where:

- oai - the scheme (which will be registered as a URI scheme)

- archive-identifier - the unique identifier of an archive (which will be registered within the Open Archives Initiative)
- record-identifier - the unique, persistent identifier of a record within the archive (the syntax of this name is archive-specific within the limitations of the URI syntax).

Sets (formally called Partitions)

The Partition concept in the original technical agreement has been retained, but has been renamed Sets. The concept allows individual records in archives to be arranged in unconstrained sets at the discretion of the archive administrator. These Sets can then be organized in a hierarchical fashion to expose the internal structure of the archive. Individual communities can make explicit agreements on the actual meaning of Sets within their communities. As such, individual communities may use Sets as a tool for selective harvesting. However, they are not meant to serve as a general tool for determining categories.

OAI Harvesting Protocol

At the heart of the technical agreements of the OAI is the metadata harvesting protocol, which provides a simple interface to transfer metadata from a data provider to a service provider. The original specifications for this interface in the Santa Fe Convention were a subset of the more expressive Dienst protocol (called the Open Archives Dienst Subset). Discussions at the meeting revealed that while the semantics of the subset service requests were generally correct, many of the artifacts of the broader Dienst protocol presented unnecessary complexities to implementers. As a result an independent OAI protocol was developed, derived from the original Open Archives Dienst Subset.

This protocol contains the following service requests:

- *Identify* – returns a self-description of the archive, containing information submitted at time of registration and other administrative information;
- *ListMetadataFormats* – returns a list of identifiers of metadata formats that are offered by the archive in general or for a particular record;
- *ListSets* – returns a structured list of sets (formally called partitions) within which records may be located;
- *ListRecords* – returns a list of record identifiers, and optionally metadata, within a specified range of dates and/or a specified Set. Flow control is achieved by the use of server-generated continuation tokens;
- *GetRecord* – returns the metadata associated with an identifier.

These service requests and their parameters are encoded into standard HTTP URIs. Responses to the requests are XML documents. This allows for simple implementation using CGI scripts or similar technology for data providers while service providers can exploit the proliferation of XML parsers to ease the harvesting of data.

Registration

It was decided that the process of registration should become more automated. Also, it was decided that the OAI should currently keep registration of compliant data providers (archives), compliant service providers and metadata formats under its own governance.

Information elements that need to be included for the registration of a data provider have been reconsidered in light of the extension of the scope of the Initiative. The existing registration via the provision of a data provider template will be replaced by:

- On-line registration of an archive identifier;
- On-line registration of the BASE-URL of the archive's OAI Protocol implementation;
- Support by the archive of the Identity verb that will expose essential information about the archive's machine interface, its policies, etc.

Registration of metadata formats – including format identifier, description of the format's semantics and the DTD of its XML transportation format -- will be automated and will include a check of the validity of the DTD.

Registration of service providers has not been discussed. A revision of the information elements required for registration will be proposed.

Acceptable Use

Acceptable use of data

The "gentlemen's agreement" between data providers and service providers, whereby

- The data providers expresses usage restrictions for data harvested from its archive;
- The service provider expresses to comply with those restrictions;

is maintained. However, an explicit distinction should be made between the harvesting of metadata -- which is the topic of the specifications -- and the harvesting of the full-content – which may become accessible via keys in the harvested metadata.

Acceptable use of the harvesting interface

Verification of the identities in the harvesting transactions is not part of the current specifications, but it is left to individual communities to use appropriate tools (such as HTTPS, TLS) if required.

A flow control mechanism built into the harvesting protocol and error messages will give archives some level of control on the usage of their harvesting interface by service providers.

Acknowledgements

Participants at the meeting

Caroline Arms – Library of Congress

Ray Dennenberg - Library of Congress

Daniel Greenstein – Digital Library Federation

Thomas Krichel – University of Surrey

Carl Lagoze – Cornell University

Xiaoming Liu – Old Dominion University

Clifford Lynch – Coalition for Networked Information

David Millman – Columbia University

Michael L. Nelson - NASA

John Ober – University of California

Thorsten Schwander – Los Alamos Laboratory

David Stuve - MIT

Robert Tansley – University of Southampton

Hussein Suleman – Virginia Tech

Simeon Warner - Los Alamos Laboratory

Herbert Van de Sompel – Cornell University

Meeting Supported by

The Digital Library Federation

The Cornell Digital Library Group

Report by

Carl Lagoze – Cornell University - <lagoze@CS.Cornell.EDU>

Hussein Suleman – Virginia Tech - <hussein@vt.edu>

Herbert Van de Sompel – Cornell University - <herbertv@cs.cornell.edu>

Dienst
Overview and Introduction

What is Dienst?

The word "Dienst" refers to a number of things:

- A conceptual architecture for distributed digital libraries.
- A protocol for service communication in that architecture.
- A software system that implements that protocol.

Conceptually, Dienst is a system for configuring a set of individual *services* running on distributed *servers* to cooperate in providing the services of a digital library. The *open architecture* of the Dienst system - exposure of the functionality through an defined protocol - makes it possible to combine Dienst services in flexible ways and augment the existing services with other *mediator services*, which build on the functionality of the existing services.

The Dienst system originated with the Computer Science Technical Reports Project, a DARPA-funded collaboration to establish a digital library of computer science technical reports.

What can Dienst be used for?

Broadly speaking, there are three ways that Dienst is used:

1. Organizations that wish to join an existing digital library built using Dienst.
2. Organizations that wish to create a new distributed digital library with Dienst.
3. Organizations that wish to undertake research in digital libraries using an existing Dienst digital library or by experimenting with the software. The modular nature of the software encourages researchers who wish explore mechanisms for enhancing existing services or using the interfaces to existing services to built other services.

***Notice*:** While the software has been built and tested according to conventional software engineering standards, we stress that it is *research software*. Organizations wishing to use it for mission critical applications should consider a thorough review of the architecture, protocol, and software for robustness and security.

What kind of resources can Dienst be used for?

The distributed Dienst software is configured to handle textual resources (documents) in a variety of formats. However, the Dienst architecture includes a sophisticated document

model that accommodates a wide variety of digital resources. Using the Dienst software for these other resources will require some programming.

How is Dienst licensed?

The Dienst protocol and software is copyrighted but is available for free and can be used and redistributed for *non-commercial* uses.

Who uses Dienst?

The Dienst protocol and software is used in a variety of existing digital libraries and projects. Some of these are listed below:

- NCSTRL, the Networked Computer Science Technical Reference Library.
- CoRR, the Computing Research Repository.
- The Open Archives Initiative.
- ETRDL, the ERCIM Technical Reference Digital Library.
- Cornell University Library Historical Math Book Collection
- Cornell University Library Making of America Collection
- Hein online Retrospective Law Journals

Who supports work on Dienst?

Dienst is a project of the CDLRG - Cornell Digital Library Research Group. Work on Dienst sponsored by the Defense Advanced Research Projects Agency (DARPA) on behalf of the Digital Libraries Initiative under Grant No. N66001-98-1-8908. Additional work on Dienst is sponsored by the National Science Foundation Digital Libraries Initiative Phase 2 Project Prism under Grant No. IIS-9817416.

More Information?

More information is available by reading the other documents available at this site:

- Architecture Summary
- Protocol Specification
- Software Overview

Also we have written a number of papers have been written about Dienst and its applications. Refer to the References section.

Send mail to info@prism.cornell.edu for more information.

References

Predicting Indexer Performance in a Distributed Digital Library, Cornell University Technical Report and draft of submission to European Digital Library Conference, May 1999.

Using Query Mediators for Distributed Searching in Federated Digital Libraries, Draft of submission to ACM DL'99, August 1999.

A Characterization Study of NCSTRL Distributed Searching, Cornell University Technical Report, January, 1999

Defining Collections in Distributed Digitial Libraries, D-Lib Magazine, November 1998.

NCSTRL: Design and Deployment of a Globally Distributed Digital Library, Draft of submission to Journal of the Society of Information Scientists (JASIS) 1999.

Making Global Digital Libraries Work: Collection Services, Connectivity Regions, and Collection Views. ACM DL'98, June 1998.

The Networked Computer Science Technical Reports Library. Cornell Computer Science Technical Report, July 1996.

Dienst: Building a Production Technical Report Server. Chapter 15 in *Advances in Digital Libraries*, Springer Verlag 1995.

Dienst: implementation reference manual. Cornell Computer Science Technical Report, May 1995.

Dienst - An Architecture for Distributed Document Libraries. Communications of the ACM, April 1995, Vol 38 No 4 page 47.

"Drop-in" publishing with the World Wide Web. 2nd Int'l WWW Conference 1994.

A protocol and server for a distributed technical report library. Cornell Computer Science Technical Report, June 1994.

Experiences and Organisation of the European Physical Society Portal of Services

Invoking International and National Societies and University Groups Worldwide

E. R. Hilf, T. Severiens
Institute for Science Networking, Oldenburg, Germany

Since 1995 the European Physical Society has offered free access web services for all kinds of information gathered by HARVEST directly from the professional physics institutions distributed around the world. Included are personal or institutional information, documents, reports, and teaching materials. Thus the authors guarantee their rights, actuality, and integrity of the information. The information is accessed by an increasing net of linked HARVEST brokers run by EPS and national societies.

Author tools are offered to simplify submission of metadata.

The Open Archives initiative now allows access to the service by other providers by wrapping the metadata set of PhysNet <www.eps.org/PhysNet/> into that of OAi.

PhysNet is a not-for-profit service for scientists. The organizational structure (see its Charter) ensures coherent cooperation of national and international societies and institutions. EPS, by its Action Committee on Publication and Scientific Communication, controls the service and its development.

EPS has responsibility for the management of the coherence of PhysNet with analogous services of adjacent fields. The coherence with adjacent learned fields is managed through agreements and through Joint Technical Workgroups as first set up with the IMU (Int. Math.Union).

In Germany the bridge to other fields is built by the IuK, the Information and Communication Initiative of the Learned Societies. The IuK and the societies of university libraries, computer centers, and media centers at universities have set up an initiative for this called DINI.

Experiences and usage of the services are explained, and future plans described.

The EPrints.org software

Robert Tansley, University of Southampton

The Open Archives Initiative is providing a framework for the extensive interoperability of archives of scholarly research literature (and potentially other digital materials too). For this initiative to become a widespread reality, the participation of large numbers of individuals and institutions is required.

In order to be able to participate, what institutions need is working, interoperable, configurable software that is freely available and easy to set up. It is this need that the EPrints software was developed to fulfill.

EPrints is a feature-rich open archive software system that is as simple to install as any normal application. It runs right "out of the box" with a comprehensive default setup that should serve most people's needs. However, it has also been designed to make it extensively and flexibly re-configurable for customized needs; almost any aspect of the archive's operation can be adapted to suit a particular requirement.

This means that the archive can be used by institutions, individuals, journals or any other organization wishing to interoperate with Open Archive services.

This adaptability is achieved by using a modular design methodology. The system is divided into two main components: The core archive component, which provides the functionality required for all open archives, and the site-specific component, providing details about exactly what is stored in the archive, how it is presented and how it may be searched. The system is supplied with a richly featured site-specific component that requires minimal changing (handled by an installation script) to set up a fully working, interoperable open archive. When updated revisions of the software become available, the core archive component can be upgraded, and the site retains its identity and data in the site-specific component.

The many aspects of the software that can be configured by an institution include:

- The types of record that can be stored in the archive, and what metadata fields to hold with each;
- The types of document file (or other data) that can be stored with each record;
- The validation checks that are performed on each incoming record, to minimize administrator effort;
- Which metadata fields should be searchable by users;
- What metadata to present records to the open archives protocol (i.e. how the internal metadata maps to the open archives metadata);
- Full control over the "look and feel" of the archive (in any language)

The software also has the following features:

- "Out of the box" Open Archives interoperability;
- Simple but very powerful submission interface;
- Local browsing and searching features;
- Inter- and intra-linking potential (papers, versions, comments, responses, citations)
- Moderation buffer for incoming submissions;
- Site maintenance can be performed by a WWW interface;
- E-Mail subscription service for users.

Of course, it is simple to add extra functionality to an archive in the site-specific component of the software.

***FURTHER INFORMATION IS AVAILABLE AT**: http://www.eprints.org/*

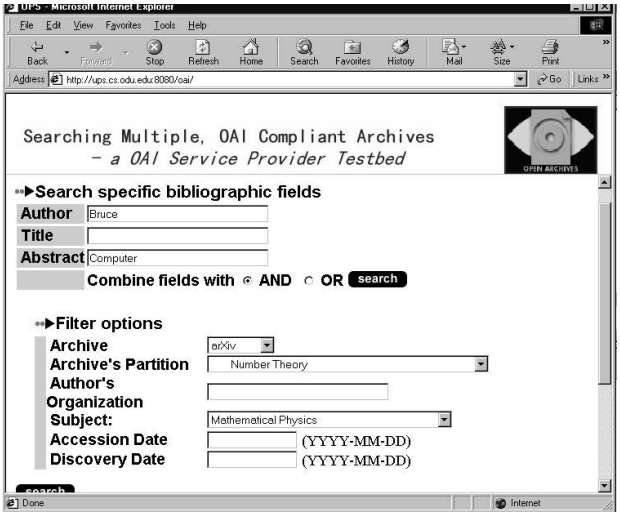
***AN ARCHIVE RUNNING THIS SOFTWARE IS AT:** http://cogprints.soton.ac.uk/*

Searching Multiple, OAI Compliant Archives: Experience in Building a Prototype and Preliminary Results

X. Liu, K. Maly, M. Zubair
Department of Computer Science,Old Dominion University, Norfolk, VA

This prototype harvests metadata from several OAI data providers in an Oracle database and implements an end-user searching service. The data providers that have been harvested are: arXiv.org, CogPrints, NCSTRL and WCR. All these archives are OAI-compliant and can be harvested using the OAI Dienst subset. For this prototype we intend to harvest all metadata of these archives (around 200K records). So far we have collected over 100K metadata records, and end–users can search/browse the harvested metadata using a single interface as shown below.

Search Interface



The interface allows users to search multiple archives by author/title/abstract. Users may filter the search or the browsing of multiple archives by archive/partition/subject/accession date/discovery date. We now briefly highlight some of the problems encountered in building such a prototype.

Problems Encountered and Proposed Solution

- OAMS and Dienst subset conformance. There are various problems, including XML syntax and encoding, case sensitivity, invalid responses, etc. We believe a verification schedule is important. We also believe all responses need strict DTD specification.
- OAI problem. We found several technical problems of OAI which hinder harvesting. Firstly, the encoding problem: we need to support ISO8859-1 or other standards. Secondly, “file-before” parameter is necessary in “list-contents” verb.
- Harvest schedule problem. Data providers have different requirements for web robots, this has to be done by human communication now. We need machine based mechanics, which can negotiate a harvesting schedule between data provider and service provider. We are working towards a channel-based model to address this problem.
- Data provider management. OAI currently uses the web page as communication between data providers and service providers. A service provider has to manually put details into software. This is not a scalable approach and a naming service is necessary.

Virginia Tech - Digital Libraries Research Laboratory

September 2000

"development of protocols and standards documents"

data archives, compliance testing, metadata formats, software implementations

Virginia Tech's Involvement in the OAi

Members of Virginia Tech's DLRL have been involved with this process from the early stages and continue to contribute towards the development of protocols and standards documents that comprise the Santa Fe Convention. Professor Edward Fox was one of the original participants in the project, representing the NDLTD (Networked Digital Library of Theses and Dissertations) project. At that first meeting Virginia Tech made a commitment to integrate NDLTD into the OAi project. Subsequent to that, the CSTC (Computer Science Teaching Center) and W3C Web Characterization Repository became two of the first OAi-compliant digital libraries. Work is also done in testing compliance of archives, defining metadata transport formats and developing client/server implementations of the protocols. Wherever possible, new and existing projects of the DLRL are being linked to the OAi to provide a proof of concept of the ubiquitous capability of the OAi's protocols and specifications.

Projects

Networked Digital Library of Theses and Dissertations – federated digital library supporting OAi protocols

Repository Explorer – technical compliance test for the OAi protocol

MARC XML DTD – XML transport format for US-MARC records

CSTC, W3C Repository, VT-Electronic Thesis and Dissertation Collection – OAi-compliant archives maintained by members of the DLRL

PERL/JAVA Software – Implementations of the OAi Protocol

Further Information

http://www.dlib.vt.edu/projects/OAi/index.html