Workshop 5

DIGITAL LIBRARIES '00

HYPERTEXT '00

Extending Interoperability of Digital Libraries:

Building on the Open Archives Initiative

3 June 2000

San Antonio, Texas, USA

Contents

Program

Note: Unless otherwise noted or decided, all general discussion will be chaired by Clifford Lynch, Director of the Coalition for Networked Information.

<table>
<tr><td><b>8:00 am – 8:30 am</b></td><td><b>Registration</b></td></tr>
<tr><td><b>8:30 am – 10:00 am</b></td><td><b>Session One – Introduction and Review</b></td></tr>
<tr><td colspan="2"><b>Introductory Remarks - "The Big Picture"</b>
<i>Clifford Lynch, Coalition for Networked Information</i> (30 minutes)

<b>The Open Archives Initiative – Where We Are Today</b>
<i>Carl Lagoze, Cornell University</i> (20 minutes)
A brief introduction to the Open Archives Initiative and progress since the Santa Fe Meeting in terms of metadata, protocol, server, and harvesting efforts.

<b>Reviews and Discussion of Implementation Issues – Part I</b>
(40 minutes)
<i>input from: Michael Nelson (Old Dominion University), Robert Tansley (Southampton University), Simeon Warner (Los Alamos National Laboratory), Hussein Suleman (Virginia Tech)</i></td></tr>
<tr><td><b>10:00 am – 10:20am</b></td><td><b>Break</b></td></tr>
<tr><td><b>10:20 am – 12:00 pm</b></td><td><b>Session Two - Review of Open Archives Initiative</b></td></tr>
<tr><td colspan="2"><b>Reviews and Discussion of Implementation Issues – Part II</b></td></tr>
<tr><td><b>12:00 pm – 1:00 pm</b></td><td><b>Lunch</b></td></tr>
<tr><td><b>1:00 pm – 2:50 pm</b></td><td><b>Session Three - Future of the Initiative I</b></td></tr>
<tr><td colspan="2"><b>General Discussion</b>
Review of morning's activities and wrapping up of any loose ends (30 minutes)

<b>The Way Forward - An introduction to the issues that need to be resolved</b>
<i>Edward Fox, Virginia Tech</i> (20 minutes)

<b>Discussion: Organization, Outreach and Dissemination</b>

<b>Discussion: Core Technical Issues</b></td></tr>
<tr><td><b>2:50 pm – 3:15 pm</b></td><td><b>Break</b></td></tr>
<tr><td><b>3:15 pm – 4:15 pm</b></td><td><b>Session Four - Future of the Initiative II</b></td></tr>
<tr><td colspan="2"><b>Discussion: Research Issues and Opportunities</b>

<b>Discussion: Scope (Context in terms of data types and domains)</b></td></tr>
<tr><td><b>4:15 pm – 5:00 pm</b></td><td><b>Session Five - Concluding Discussion</b></td></tr>
</table>

List of Participants

1. Allard, Suzie - University of Kentucky
2. Allen, Bob - University of Maryland
3. Belton, Keith - SOLINET
4. Blixrud, Julia C. - Association of Research Libraries & Scholarly Publishing and Academic Resources Coalition
5. Bollacker, Kurt - The Internet Archive
6. Breeding, Marshall - Vanderbilt University
7. Carr, Les - University of Southampton
8. Celeste, Eric - MIT Libraries
9. Combs, Jody - Vanderbilt University
10. Fox, Edward A. - Virginia Tech & Networked Digital Library of Theses and Dissertations
11. Freire, Nuno - INESC, Portugal
12. French, Jim - University of Virginia
13. Ginsparg, Paul - Los Alamos National Laboratory & arXiv.org
14. Gold, Anna - University of California - San Diego
15. Gonçalves, Marcos André - Virginia Tech & Networked Digital Library of Theses and Dissertations
16. Hall, Wendy - University of Southampton
17. Harnad, Stevan - University of Southampton & CogPrints
18. Hellman, Eric - Openly Informatics Inc.
19. Hey, Jessie M. N. - University of Southampton
20. Hitchcock, Steve - University of Southampton
21. Jiao, Zhuoan - University of Southampton
22. Lagoze, Carl - Cornell University
23. Leng, Yin - Middlesex University, UK
24. Levy, David
25. Luce, Rick - Los Alamos National Laboratory
26. Lynch, Clifford - Coalition for Networked Information
27. Maly, Kurt - Old Dominion University
28. McFarland, Mark - University of Texas, Austin
29. Nelson, Michael - NASA Langley
30. Noreault, Terry - Online Computer Library Center (OCLC)
31. Nowell, Lucy T. - Pacific Northwest National Laboratory
32. Ober, John - California Digital Library
33. Rusch-Feja, Diann - Max Planck Institute for Human Development, Germany
34. Sánchez, J. Alfredo - Universidad de las Américas, Mexico
35. Schnase, John L. - NASA - Earth and Space Data Computing Division
36. Sugimoto, Shigeo - University of Library Information Science, Japan
37. Suleman, Hussein - Virginia Tech
38. Tansley, Robert - University of Southampton
39. Van de Velde, Eric - Caltech
40. Warner, Simeon - Los Alamos National Laboratory & arXiv.org
41. Wayland, Ross - University of Virginia
42. Weiss, Ken - University of California
43. Wesley, Rebecca - Stanford University - Lane Medical Library
44. Williamsen, Julie - Brigham Young University

Position Statements

Allard, Suzie

Expanding the OAi Mission

(Topic Area: Organization, Outreach and Dissemination)

Expanding and assuring longevity of the OAi mission are important "next steps", and I believe there are two issues to address:

1. Creating a framework that can support the OAi scholarly organization and its mission.

2. Encouraging participation in OAi from a diverse range of disciplines and universities.

My comments are offered to provide a basis for conversation, but do not elaborate fully on each of the points.

First, in the broadest terms I define any scholarly organization's mission as working to enhance knowledge creation through six mechanisms: forum, multi-disciplinarity, resource sharing, perspective, sharing and connectivity. OAi touches on many of these mechanisms, and this should be conveyed to potential participants.

FORUM: creates "communication space" for interaction between scholars who are pursuing similar interests. The members of a scholarly organization are organized around a specific purpose. However each of the members may be tackling this purpose from different perspectives and it is likely that members represent diverse disciplines. When individuals enter into a collaborative project or cooperative venture, a new communication space is created that allows members to participate in discourse.

MULTI-DISCIPLINARITY: encourages communication across disciplinary boundaries, thus accelerating maturation of ideas. The scholarly organization, particularly because it is interdisciplinary, brings a unique set of intellectual capital to the task. Each member brings their own tacit knowledge and skills as well as the cultural knowledge of their discipline. Currently much of the scholarly community only publishes within its own discipline; therefore work is frequently only seen by peers (Pierce, 1990; Siow, 1995).

RESOURCE POWER: leverages resources from many different institutions. Scholarly organizations connect institutions, and allow the group as a whole to benefit from resources that may be held by one site. This increases knowledge creation by allowing ideas and concepts to be developed and tested without requiring redundant investments in resources.

PERSPECTIVE: introduces regional, cultural, and institutional perspectives to the process. By assembling a large number of members and creating connectivity between them, it is likely that differences in location or culture can be identified, discussed and appropriate action can be taken. The differences may be assimilated into the design of the organization or they may be only noted and then ignored.

SHARING: encourages the creation of new knowledge through sharing of existing knowledge. Scholarly organizations are enacted so that existing knowledge is shared. As noted before, sharing knowledge leads to greater powers of knowledge creation.

CONNECTIVITY: shortens time frame for communication. The scholarly organization establishes communication paths, which encourage members to communicate information directly without an intermediary. While some of the communication will be management oriented, the larger part will be concerning the exchange and interpretation of information. This opens up an extremely fluid conduit for bi-directional information communication which reduces the information interval between knowledge creator and knowledge user.

To create a management framework and express the OAi concept to potential participants, these are some of the steps I think we need to discuss.

1. Establish vision/goals: The goal of the organization must be clearly stated so that it can be easily understood by all involved. As Senge (1995) points out when speaking about shared vision, the scholarly organization relies on members being genuinely committed to the task/cause since that is the only motivation that can elicit the desired results and there is no mechanism for creating compliance. What needs to be communicated to all members is a well worded, brief statement of the purpose of the organization. It must allow for the different motivations leading members to participate, but it must also provide a concrete reason for the organization's existence.

2. Define roles and products: The whole consortium must know what is being tackled and the specific responsibilities associated with becoming involved.

3. Determine communication channels: Communication can make or break the scholarly organization. Whether using traditional or progressive communication methods, the channels must be established early, and members need to be encouraged to use them. These channels will facilitate all aspects of information sharing and knowledge creation. They have the capacity to remove the barriers that stifle a learning organization.

4. Establish ontology, categorizations and vocabulary: Fulfilling the purpose of creating new knowledge is very difficult if the members of the organization are not conversing in terminology that all understand. Therefore, the organization and representation of knowledge becomes an important factor. While this predominately refers to content, it also is important in terms of organizational structure. Understanding must exist between all members, however it must always be kept in mind that there are often different value systems at work.

5. Prepare for perspectives: The organization should be designed to address the different perspectives brought to the project by the institutional, disciplinary and national backgrounds of the members. For example, in a survey of 430 corporate executives, 54% said that culture (corporate) was the current biggest impediment to knowledge transfer (The Ernst & Young Center, 1999). Therefore there is reason to believe that this socially embedded characteristic may also impact knowledge on a university or disciplinary level.

6. Create structure for management: One characteristic of the scholarly organization is that it is boundaryless, that is, it attempts to dispel boundaries created by hierarchy, discipline and geography. However, even with the horizontal nature of the organization there needs to be some sort of structure that coordinates the member's efforts. Ideally, the structure of the organization itself will create a checks and balance system that will aid in the management of the project. However, it appears that it is necessary for a small group to act as a steering committee to coordinate the OAi activities.

7. Create system of evaluation: This is related to initiative management, however it is important enough that it should be considered separately. We need to have a method to access the organization in terms of goals, roles, products and outcomes, so that OAi can reflect the changing scholarly landscape.

Suzie Allard
University of Kentucky (859)257-8876
College of Communications and Information Studies
502 King Library South, Lexington, KY 40506-0039
e-mail: slalla0@pop.uky.edu

Allen, Bob

While the Open Archive Initiative has focused on standards for text e-print services, I believe that similar services should be developed for multimedia content. For instance, collections of videotapes of lectures and presentations at universities could be developed much the same way that technical reports are collected.

Indeed, a Multimedia Open Archives Initiative could be used for archives of all sorts of multimedia educational materials. The Maryland Electronic Learning Community (http://www.learn.umd.edu) has carefully constructed a multimedia digital library, but many of the most popular items are activities and local resources that the teachers have entered themselves through a quick indexing tool (Semple, Allen, & Rose, Edmedia 2000). We could envision a national Open Archive by which teachers could exchange resources.

Bollacker, Kurt

From the perspective of technical authors, the Open Archives initiative has the noble goal of bringing order to the submission process of their papers and metadata to electronic archives. However, there is also much interest (and hype) in using the Web directly as a publication venue. Some authors post a paper to their Web site, and don't bother submitting to an e-print archive. Some organization may have institutional support for Web publishing but not e-print submission. Thus, many papers exist on the Web but may not ever make it into an e-print archive, and I believe this condition will continue for a long time. If OAi standards and systems consider only manual submission of documents, then a large base of potential growth may be missed.

One approach to capitalizing on this potential opportunity is to consider the use of Web crawling technology as a complement to manual submission. If authors place papers on their home pages, automated archiving agents may be able to find, download, and parse the papers to retrieve the appropriate metadata and source documents. This approach has several advantages over exclusively manual submission:

1. It is easier for some authors. They simply "post and forget" their paper without worrying about other venues for electronic publication.

2. Many papers are published on the Web, rivaling even large existing e-print archives. Thus, automated methods supplementing manual inclusion may result in larger and more complete archives.

3. It is retroactive. Papers published on the Web in the past can be automatically included in the future.

Despite the benefits of such an approach, there are disadvantages of using the Web as a source of content for e-print archives:

1. Due to the non-standard form and content of papers, it may be difficult or impossible to find some papers or to parse and extract from them the desired metadata and content.

2. There is uncertain provenance of documents, since it would be easy for one person/organization to publish the paper of another on the Web. No easy verification of metadata is possible.

3. The completeness (WRT a particular author/organization) is not guaranteed because automated Web crawlers may miss some items.

An existing example of an automated archive, ResearchIndex, contains about 270,000 papers, primarily from the field of computer science. This is quite large compared to many existing e-print-type archives requiring manual submission. While the average quality of the information extracted about a paper in this archive may be lower than a fully manual e-print archive, it does achieve a high level of usefulness, which cannot be ignored.

I propose that the OAi should be "Web robot aware". I suspect that encouraging "Web friendly" standards and protocols would not be a large burden, and would have major benefits over time. To do this we simply need to keep in mind that at some layers, there could be either a person or a piece of software interacting with an archive. A few of the issues arising from this idea include:

1. Omission of "required" information - There should be tolerance to missing data in even some non-optional metadata fields.

2. Standardization of document/metadata submission protocols - If the submission process is standardized at some technical layer, then software for automated submission may be able to co-exist nicely with manual submission.

3. The need for duplicate detection - This includes automatically identifying different versions of the same document or documents with trivially different metadata.

Despite the higher initial overhead, I believe that the hybrid approach of considering both manual submission and automated crawling in setting standards and building e-print archives will result in higher completeness and quality in the long run.

Fox, Edward A.

Background - NDLTD, NUDL, NSDL:

Since the mid-1980s we have been working in the confluence of technologies and practices related to production, distribution, management, and use of information. Electronic publishing, scholarly communication, multimedia, information retrieval, hypertext, networked information, user interfaces, digital libraries, and other areas have contributed tools, systems, techniques, and services that are leading to rapid change and innovation. Three efforts in particular illustrate this, and have been connected with many of the efforts at Virginia Tech (see also statements by Gonçalves and Suleman).

NDLTD, www.ndltd.org, Networked Digital Library of Theses and Dissertations, is a rapidly growing federation of universities, libraries, and other institutions interesting in author-submission of scholarly research, management of submissions by local and support organizations, and appropriate access to thousands of these large (multimedia) works. Based on work that began in 1987, there are well over 100 members of NDLTD, with statewide (Ohio), regional (Catalunya), national (Australia, Germany, Portugal), and international (ISTEC, supporting Ibero-America - also aided by OAS and UNESCO) consortia collaborating. This free and open federation shares training materials, software, tools, metadata, and electronic documents using digital libraries. There will be a single Open Archives collection for union access with content in many languages, using the MARIAN system to support harvesting, gateway access, federated search, browsing, and other services. Universities and other members also will be encouraged to have Santa Fe compliant archives that support a Dublin-Core based metadata standard being developed for electronic theses and dissertations (ETDs).

To extend the work of NDLTD toward other university information, the Networked University Digital Library (NUDL) was proposed early in 1999. NSF is supporting a small project to promote this development, focusing on collaboration with the Dissertationen Online effort in Germany, funded by DFG. Some of the content to be considered beyond ETDs is computer science courseware. The Computer Science Teaching Center, CSTC, www.cstc.org, hosted by Virginia Tech, collects such peer reviewed works. As of early 2001, some of the works in CSTC will also become a part of the ACM Digital Library, after passing additional, more stringent reviewing as required by the new ACM Journal of Educational Resources in Computing (JERIC). Entries in JERIC will have their full content in the ACM DL, but metadata for them will be in both the ACM DL, and in somewhat richer (more detailed, using IMS metadata elements) form in CSTC. CSTC is one of the efforts that have been supported by NSF DUE as part of the move toward the National SMETE (Science, Mathematics, Engineering, and Technology Education) Digital Library (http://www.ehr.nsf.gov/EHR/DUE/programs/nsdl/). We believe that most if not all of NSDL should be accessible through Santa Fe compliant open archives.

Institutions in Addition to Disciplines:

The work with NDLTD, as well as with NUDL, has emphasized that universities and other institutions should engage in development of digital libraries. These can become key parts of the institutional memory, with students, faculty, staff, and others contributing locally developed works directly into a digital library managed by the institution (e.g., by its library). This type of effort complements the development of open archives that are built around a particular topical or disciplinary focus. There are many different issues related to building digital libraries with this emphasis - see various papers about this at http://www.ndltd.org/pubs/.

Educational Activities and Educational Content:

If open archives are developed by universities, it is logical to connect the educational mission of a university with the research and scholarly communication missions of that institution. Helping students learn to produce electronic documents is valuable as a way to prepare them for the Information Age. It allows them to more easily produce documents to be submitted to conferences and funding agencies, for example, or to ship through email or over the Web. It provides a framework for students to learn about important issues that relate to communication of research results: intellectual property rights, multimedia/hypermedia technologies, working with thesauri, selecting keywords and descriptors, making preservation easier by following standards, employing descriptive markup languages, developing effective queries and routing profiles, and learning to write effectively. At this point in their careers, students can benefit from training aimed to help them learn to communicate more effectively, and have grounding in the key issues, facilitating future lifelong learning as related technologies develop. At Virginia Tech, students have been required to submit their theses and dissertations electronically since 1997, resulting in much larger numbers of submissions than would occur if submission was optional. Making a requirement is reasonable since the process is easy and the educational benefits are clear.

Technical Interests - Efficient and Effective Services:

OAI faces many challenges. Virginia Tech is interested in exploring many of the technical and other challenges that must needs be faced. We offer storage support on VT-PetaPlex-1, which has 2.5 terabytes of disk space. We offer to partner in mirroring activities, since our site has very high speed connectivity to Internet2. We will develop information retrieval (including full-text and multimedia), probabilistic filtering, harvesting, gateway, federated search, and interoperability testing services. We will help with support of MARC representations and interchange. Our aim is to help with the scaling of efficient and effective services.

Dissemination and Growth:

Based on work with NDLTD, we offer to assist with the dissemination of information about OAi and to support its growth with educational and other resources. Our approach will be to encourage universities to join NDLTD and gradually expand their archives into other types, leading to growth of NUDL. The next annual meeting about ETDs, which we hope will have about 400 people and which will take place Spring 2001 in Pasadena, will be an important venue to share information about OAi.

Ginsparg, Paul and Warner, Simeon

arXiv

Organization, Outreach and Dissemination

The most pressing need is for services that use the 7 archives currently or soon-to-be OA compliant (see http://www.openarchives.org/sfc/sfc_archives.htm). The OAi will not have realized its goals until such services demonstrate how low-level interoperability can lead to a multi-disciplinary digital library. Such services are also the best way to test the server implementations, and a good way to refine the subset Dienst specification.

The OAi's primary focus should be the technical agenda, and should aim to be as inclusive as possible. It can move beyond the original conception of preprint or "author self-archived" documents to encompass as many of the needs of the primary library community as possible.

Core Technical Issues

The OAi might also agree on a set of rules/standards for full-data availability and perhaps create a second compliance level. There could be ‘OA (meta-data) compliance’ and ‘OA (full-data) compliance’. We see no harm in agreeing on an additional OAMS field for full-data recovery; this would be optional for meta-data compliance and compulsory for full-data compliance.

As we have argued before, we do not support making unrestricted access to full-data (even if only to certain "compliant service providers") a requirement of OA compliance for data-providers. Our reasons include:

1) Experience with arXiv convinces us that we do not want to agree to allow unlimited robotic access to the full-data on arXiv.

2) We do not want to exclude archives that either do not have full-data or do not give it away.

3) We do not want to discourage data-providers from at least complying with the current OA requirements by `raising the bar' in this way.

Gonçalves, Marcos André

Position Statement for NDLTD and Marian Related Efforts
mgoncalv@vt.edu

NDLTD (Networked Digital Library of Theses and Dissertations) has been registered as an archive supporting the Santa Fe convention since the beginning of those efforts. NDLTD is an international federation of universities, libraries, and other supporting institutions focused on efforts related to theses and dissertations, including their content. Many universities run their own programs and services, but there also are consortia activities at the state (OhioLINK), regional (Catalunya), and national (Australia, Germany, Portugal) levels. Each activity is independently and autonomously managed, having just to agree on common goals and methods, as well as some basic interoperability issues, such as to provide URNs and metadata records for all stored works.

Starting at the beginning of 2000, NSF began to fund a project that liaises with a similar project funded by DFG (in Germany), for developing a multilingual federated search system to support the Networked University Digital Library (NUDL) initiative, in which NDLTD is included. We rapidly considered to extend the objectives of that project to implement the framework presented in the Santa Fe convention, due to the emergence of the Open Archives initiative as a reasonable partial solution for interoperability problems. Our efforts are based on our MARIAN system, a knowledge based digital library and information retrieval system. The work of making MARIAN compliant with the Santa Fe convention is concentrated in three different fronts:

1) make MARIAN serve as a middleware layer, able to deal with the heterogeneity of many specific sources and protocols, including OAi sources;

2) implement the Dienst OAi interface over MARIAN using a lightweight Java based implementation;

3) develop an XML transportation format for MARC records.

Considering the first objective, we have written several mediator/wrapper modules to integrate a set of heterogeneous sources inside MARIAN. We are now able to handle the complete Dienst and the OAi subset protocols, plus Z39.50 and Harvest sources. Due to the problems of dealing with so many different systems, NDLTD is now encouraging all members to implement the OAi framework. At the moment, we have performed experiments with four archives in our collection:

1) Two for the Virginia Tech ETD collection (2105 records), which include respectively MARC and XML-OAMS based data;

2) The German PhysDoc collection (1618 records) in SOIF format;

3) MIT-ETD collection (4000 records) in RFC1802 format from MIT using the complete Dienst protocol and;

4) West Virginia-ETD (675 records) collection, using SOIF-HMTL incorporated through the use of the Harvest System and a particular HTML wrapper.

Difficulties we found during our experiments include:

1) how to integrate all the heterogeneous data in a reasonable way inside MARIAN (e.g., applying fusion and object-oriented techniques);

2) how to export/map those specific metadata to our XML-MARC format; and

3) how to deal with aspects of data quality.

We have developed an architecture for integration that we think is reasonable and extensible and we briefly will make MARIAN to produce coherent ranked lists for all collections. We are now discussing semantic aspects related to the different metadata standards, and implementation aspects related to mapping between those metadata standards and supporting non-registered ones, like SOIF-Harvest. Another current problem is related to dealing with sources that use protocols unable to export complete metadata sets, like many Z39.50 sources or limited CGI-based sources.

Secondly, we have also started our lightweight Java implementation for the OAi Dienst protocol. Our current implementation focuses on the MARIAN paradigm, but we have plans to extend it to other data models (e.g., relational databases).

Finally, the existing metadata standard for many libraries is MARC records. MARC records have their own native transport format ("tape format"), but it requires specialized parsers and makes use of some fairly arcane conventions. We have undertaken to create an XML DTD to support wider distribution of MARC records within the OA community.

Existing questions include whether to more closely control field IDs and labels, and how best to convert ANSEL characters to XML entities. Our Java MarcRecord object can now produce MARC Tape Format as well as XML. The program has been tested on over 4,000 MARC records, moving from Tape Format to XML and back to Tape Format without losing a character. The MarcDocument object can also produce something approximating OAMS metadata records.

Harnad, Stevan

My contribution to OAi is nontechnical. I have an multidisciplinary archive (CogPrints) that has just been redesigned as a Santa-Fe-compliant Open Archive by Robert Tansley, who is also making a generic version of that Open-Archiving software (EPrints) so any University or Institution can immediately set up Santa-Fe-compliant, interoperable Open Archives for all of their disciplines. I am also part of the Open Citation Linking project (OpCit), in which Les Carr and Zhuoan Jiao are citation-linking the full-text contents of Open Archives. I also have online archives for two journals I edit, one a paper journal (BBS) and one an online journal (Psycoloquy), both to be transformed shortly into Open Archives.

My objective is to help facilitate and promote author self-archiving of both the unrefereed and the refereed research literature in Open Archives in all disciplines. The generic open archive software and the citation linking service are intended to hasten this process.

http://cogprints.soton.ac.uk/
http://www.eprints.org/
http://opcit.eprints.org/
http://www.princeton.edu/~harnad/bbs/
http://www.princeton.edu/~harnad/psyc.html

Stevan Harnad
Professor of Cognitive Science
Department of Electronics and
Computer Science
University of Southampton
Highfield, Southampton
SO17 1BJ UNITED KINGDOM

harnad@cogsci.soton.ac.uk
harnad@princeton.edu
phone: +44 23-80 592-582
fax: +44 23-80 592-865
http://www.cogsci.soton.ac.uk/~harnad/
http://www.princeton.edu/~harnad/

Hellman, Eric

President
Openly Informatics, Inc.

I'll be attending for a few reasons:

1. Openly's LinkBaton System presents an ideal mechanism for linking to and from open archives. LinkBatons are links that learn. They can learn about a user's library and about her access to various information sources, making it possible for the links to adapt to a user's environment. LinkBatons can link from anything to anything without requiring either party to know about each other. http://my.linkbaton.com/

2. To make links to open archives, via LinkBatons or via conventional links, the Open Archive Initiative must develop robust, open and free authorities which identify every participating archive. Openly has the experience and motivation to host such an authority or advise others in developing such an authority.

Hey, Jessie M. N.

I would like to commend the tremendous work done so far within the Open Archives Initiative and emphasise the usefulness of the prototype in exploring the practicalities. Having worked with end users in Science and Technology for many years and now staff and students within the Arts and Humanities I would like to reiterate the needs of the broader based end user. I would like to ensure that we continue to provide simplicity in submission for the author and always provide at least one simple format of full text (which does not require a practising scientist’s knowledge to retrieve). Confidence is important in instilling regular usage. We are now serving more part time and distance learners in a broader variety of disciplines who, with the advent of more journals in electronic form, are genuinely delighted to find a fulltext alternative immediately. However, training, as, for example, instituted for the NDLTD, is likely to be a necessity as processes change.

The scaling up requires a sound framework of subject organisation combining the most cost-effective balance of traditional practices and modern information retrieval techniques. My colleague Robert Tansley (OpCit project) has raised some concerns about partitions. However, being able to manipulate large numbers of results effectively may help to smooth over the idiosyncrasies of individual databases when placed together. The prototype is a practical way of exploring these ideas.

Jessie M.N. Hey
Project MALIBU – University of Southampton Libraries
(Managing the hybrid Library for the Benefit of Users)
IAM Research Group

University of Southampton
jmnh@ecs.soton.ac.uk

Hitchcock, Steve

Topic: Core technical issues

A Storage Architecture for Full-Text Access to Open Archives

One goal of the Open Citation (OpCit) Project (http://opcit.eprints.org/) is to integrate and develop software for reference linking in large open archives. To create cross-archive services that add value (e.g., linking, indexing, etc.), OAi data providers need to support a harvesting interface that allows OAi service providers to periodically poll the archives and access the full-text data relevant to their end-user services. The need for this capability has become apparent both from our own experience in developing a reference linking service and from the work of others (e.g., the UPS Prototype Project).

The current OAi framework does not define such an interface. Eprint archives compliant with the Santa Fe Convention only provide a means for collecting limited metadata which are not rich enough for building services, such as a reference linking service, on top of them.

To solve this problem we propose extending the current OAi framework in the following ways:

(1) data providers provide a machine interface for service providers to access the full-text content of the archive data; a copy of the archive data could be stored separately from the end-user interface for this purpose;

(2) authorised (i.e., Santa Fe-compliant) service providers are allowed to retrieve the full content from this machine interface;

(3) extend the OAMS to include information (e.g., URL) for retrieving the full text of a document (the present Display ID metadata do not serve well for automatic full-text retrieval);

A related issue: how are services from service providers to be integrated into the data providers' end-user environment? When users visit a data provider site, how do they know that third-party value-added services are available?

This position statement is supported by: Les Carr, Zhuoan Jiao, Steve Hitchcock and Stevan Harnad
for the Open Citation Project.

Lagoze, Carl

At this time the entire concept of the Open Archives initiative, both the technical and organizational agreements, is largely untested. We need to use the opportunity of the San Antonio workshop to thoroughly review and solidify existing agreements rather than hastily moving ahead. From the technical perspective this means picking apart the components of the Santa Fe convention and seeing if they make sense:

1. Identifiers – Is our model for unique identifiers for archives and records within them correct? Does it conform to other URI syntaxes?

2. Metadata – Do we have the right elements in the OAMS? Is the notion of a core metadata set correct? Have we properly considered conformance with other standards?

3. Protocol – Does the protocol offer sufficient functionality? Is the documentation clear for implementers?

From an organizational perspective, there are a number of key questions.

1. Scope – What is the scope of participations: Eprints? Other scholarly materials? Conventional Publishers?

2. Service Expectations – What expectations for stability and service do we have from participants? Federations such as NCSTRL have suffered from participants that offer poor levels of service.

3. Funding – What is the long-term funding model for both participating archives and for the OAi itself? Do we need some form of stable organization to act as a maintenance and certification authority or can we just put the agreements out there on a web page and let them propagate?

We need to resolve these issues before building more complex agreements and structures. The temptation to expand and complexify without building firm and verified foundations has been a problem for other similar initiatives.

Nelson, Michael

The NASA Langley Research Center / Old Dominion University Position on the Open Archives Initiative

Michael L. Nelson <m.l.nelson@larc.nasa.gov>
Kurt Maly <maly@cs.odu.edu>
Mohammad Zubair <zubair@cs.odu.edu>

The Open Archives initiative (OAi) is the most important short-range interoperability effort that we are aware of in the digital library (DL) community. OAi has the opportunity to break the currently vertically integrated system of information providers and service providers. If a sea of open archives unfolds, critical mass can be achieved and competitive DL services can arise that provide value-added services and customer-tailored services from a common corpus.

However, before this vision can be achieved, a number of issues must be addressed. Among these issues, we include:

1. Terms & Conditions (T&C) -- Currently, the OAi protocol makes no provision for T&C. It should be noted that not all information providers will wish to indiscriminately share their information with everyone. However, the OAi protocol and supporting technologies should be used for all applications, including those where non-trivial T&C apply.

2. Metadata -- The Open Archive Metadata Set (OAMS) is unnecessary. We do not need another metadata format. We should use Dublin Core (DC) if we want a simple, all-purpose format. Domain-specific applications will likely traffic in their own metadata formats, in their own encodings. Resources should not be expended to reinvent DC.

3. Testbed Systems -- Critical mass is needed to have OAi fully in use. We should always err on the side of simplicity with a nod to ease of implementation and adoption. Few have the luxury of creating entirely new DLs; many DLs have evolved from earlier DLs. Information providers are likely to be very invested in their current infrastructure, and if small, incremental additions are not part of OAi, it will sink under its own weight. We need to have working implementations, preferably multiple implementations to underscore that we are advocating a protocol, not a software product. We need to have wide early adoption of OAi implementations, both by information providers and service providers. We need to have multiple prototype DLs, with different service levels, that harvest from OAi compliant archives. We believe to encourage adoption and focus attention on OAi, a TREC-like “competition" should be sponsored for the upcoming ACM/IEEE joint DL conference. There should be guidelines for how information providers can participate, and guidelines for how service providers can participate.

Nowell, Lucy T.

The InfoViz team at Pacific Northwest National Laboratory* has developed a variety of tools for the synthesis, analysis, and visualization of information. Information visualization systems enable users to take advantage of the human perceptual system to build understanding of the information space. At their best, information visualization systems support exploration and discovery, as well as formation and testing of hypotheses about the collection.

Additional details are available at http://www.pnl.gov/infoviz

Users of information visualization systems often initially approach a collection with minimal information about it and about what kind of exploration is needed.

Thus, we are concerned that new architectures support users in working with a wide variety of metadata and document formats, and with potentially vast, dynamic collections of information objects.

*Pacific Northwest National Laboratory is managed by Battelle for the U. S. Department of Energy under contract number DE-AC06-76RL0 1830.

Ober, John

University of California/California Digital Library

The following progress report indicates the focus and scope of the CDL's eScholarship initiatives, which include a core commitment to e-print services and their interoperability. CDL representatives participated in the OAI Santa Fe discussions and are interested in contributing to the continuing discussions, research, technical developments, and partnerships that will best serve innovation in scholarly communication and in interoperable e-print repositories.

Electronic Scholarship Initiative – Progress 1999-2000

BACKGROUND

Founded in October 1997 by President Richard C. Atkinson, and led by University Librarian Richard E. Lucier, the California Digital Library (CDL) has established itself as the digital "co-library" of the University of California. Organizationally housed at the UC Office of the President, the CDL operates in close collaboration with all UC campuses and many of its management and operations staff are campus-located.

From its planning and inception, the CDL has been viewed as a comprehensive system for the management of scholarly information, from its inception through its organization, dissemination, and use. Many of the California Digital Library's activities to-date have been focused appropriately on building, managing, and providing access to shared scholarly collections. However, as part of its mission, CDL activities necessarily include applying appropriate digital technologies to influence and support innovations in scholarly communication, and the development of sustainable models in a digital environment.

To this end, the CDL has initiated its electronic scholarship ("eScholarship") program, the overall goal of which is the development of an infrastructure for digitally-based scholarly communication that:

· Facilitates the expressed mutual interests of the University, its faculty, and the broader scholarly community;

· Leverages the formidable capabilities and strengths of the University of California in order to provide effective national leadership in this area; and

· Supports and extends experimental reconfigurations of the components of scholarly communication by communities of scholars themselves.

eScholarship Initiatives: A Beginning

The first eScholarship initiative is University ePub, a program that focuses on using digital technologies to create UC supported electronic publications and support services. The three-year development plan for University ePub will result in an electronic print (e-print) server, the migration or establishment of new electronic journals drawn from the server, and tools for submission, access, certification and other scholarly activities.

A second initiative focuses on primary source publications. This program will bring together digitized primary photographs, art, sound recordings, and other primary source materials from the rich collections of the Online Archive of California (OAC) with scholars and editors to produce new digital books. The primary source publications program will allow novel forms of humanities and social science scholarship based upon primary source materials, as well as provide broader access and relevance to additional audiences, for example, digitized artifacts and commentary about California history for the K-12 community.

SCHOLARS: THE CDL'S STRATEGIC PARTNERS IN eSCHOLARSHIP

· Physics, Mathematics, and Computer Science: The CDL has mirrored the arXiv e-print server developed at Los Alamos National Laboratories and entered into a long-term collaboration with Paul Ginsparg, its chief architect and scientist, to experiment with a broad range of enhancements. These key scientific communities and their use of the arXiv server represent some of the most innovative and successful experiments to date in scholarly communication.

· International and Area Studies: Led by scholars from over 20 centers and organized research units from eight UC campuses, successful planning has led to the formation of an advisory and editorial board which will work with the CDL to establish a University ePub digital working papers repository and experiment with peer-reviewed publications to be drawn from it.

· Electronic Cultural Atlas Initiative (ECAI): eScholarship is the focal point for collaboration that will provide persistent access to, and innovative publications drawn from this acclaimed international effort - led by UC Berkeley Professor Lewis Lancaster - to produce standards-based humanities and social science datasets.

· Archaeology: Led by the Cotsen Institute for Archaeology at UCLA, an agreement in principle has been reached for support and development of a new digital monograph series, a digital journal, K-12 interactive websites, and a repository of field data from which to produce additional scholarship. It is expected that the first 2-4 monographs, with a new CDL/UCLA imprint, will be published within 12-24 months.

· Dermatology: Migration to the CDL in Spring 2000 of the Dermatology Online Journal, established at UC Davis, represents support of an existing UC faculty-led innovation and opportunities for further experiments in supporting peer-reviewed scholarship and access to rich supplementary materials, such as the images and data sets associated with clinical medicine. An international editorial board has been assembled for this effort.

· Tobacco Control, Pharmaceutical Chemistry, Environmental Science: Leading UC-based scholars in these disciplines have identified their readiness for experimentation and desire to explore possible experiments associated with University ePub. Discussions are continuing with these individuals and related societies.

· UC Press: As a current CDL partner, a key participant in discussions about International and Area Studies digital publications and Online Archive of California Publications, the UC Press is poised to contribute editorial, marketing, and other publications expertise to the eScholarship organizational infrastructure.

· Scholarly Societies: From the outset eScholarship has envisioned strong partnerships with societies to identify new scholarly products and contribute to editorial and advisory processes. Conversations with several societies, including the American Physical Society, the American Chemical Society, and the American Association of Pharmaceutical Sciences have confirmed their interest. Active partnerships with archaeological, dermatological, and east Asian studies societies are being pursued under specific projects outlined above.

· Foundations and Granting Agencies: The University ePub initiative has recently received support from the Scholarly Publishing and Academic Resources coalition (SPARC). The $167,000 grant is one of three awarded in SPARC's competitive Scientific Communities Initiative, designed to spur digital science publishing ventures based in academe. Proposals to examine business models, copyright, and fostering innovation are being prepared for Summer 2000 submission to the Mellon Foundation, which has ongoing interest in scholarly communication and UC's eScholarship activities.

TECHNOLOGY AND INNOVATIONS

· Electronic scholarship repositories: The January 2000 establishment of a development mirror of the LANL arXiv e-print server has seeded the development of an extensible technology architecture and the initial design of support services for submission, discovery, and cross-linking of scholarship.

· Interoperability standards: eScholarship principals contributed to international discussions for the interoperability of e-print repositories, which have led to a draft set of interoperability standards called the Open Archives Initiative. University ePub will be an early implementer of the standards.

· Establishing and migrating electronic journals: Migration of the Dermatology Online Journal to University ePub is scheduled for January 2001; establishing e-Xcavation, a new digital Archaeology journal, is tentatively planned for 2001. In all cases, author and reader support services and linking to supplementary materials are key components that will be identified and specified by scholars themselves through editorial boards and working partners.

· Digital books: With advice and assistance from the UC Press, editorial processes for the assembly and production of digital "books" from the Online Archive of California and International and Area Studies scholarship are in planning, and a joint project team has been assembled for this purpose.

EXPERTISE

· Advisory and editorial processes: Assembling an overarching steering committee for eScholarship initiatives, to be appointed at the Presidential level, is planned for late 2000. Community-based editorial and advisory boards are planned for each supported community; the International and Area Studies board was appointed in February 2000 and the Dermatology board was formed in March 2000.

· Key expertise: UC faculty experts in copyright, business and economic planning, and advanced technologies have been contacted and are expected to serve on the steering committee or as consulting partners to eScholarship communities and initiatives.

· Assembling staff: A Director of Scholarly Communication Initiatives and a Web and Services Design Coordinator have been named following an international recruiting effort. Other core staff members are under recruitment. eScholarship is already following the successful CDL strategy of employing campus-based staff on a rotational basis, to ensure cross-campus collaboration and perspective.

Rusch-Feja, Diann

The Max Planck Society (http://www.mpg.de) consists of ca. 80 pure research institutes all over Germany which are dedicated to high level, interdisciplinary research. There are currently plans for establishing a Center for Information Management which will include service functions and coordinate efforts for innovation and information research. Among other goals, it is felt that the individual institutes of the Max Planck Society should improve visibility of their own research results and publications in the form of increased participation in preprint / archive servers, and if feasible in areas where no such server is established, also build up corresponding services in that field.

The International University Bremen is similarly interested in participating in those areas which are earmarked for scholarly and professional leadership. This University is currently in its initial developmental phases, but promises to provide a key site for potential leaders in information management. Its unique development will be characterized by strategic information alliances and an almost completely electronic information environment. New structures for information provision and information technology services, innovation in educational and curriculum development in the context of this environment will be monitored for insights as to how electronic information resources and technology specifically change higher education goals, student behavior and professional qualification, etc.

I will be representing both institutions at the OAI meeting on Saturday.

Sugimoto, Shigeo

Univ. of Library and Information Science

Tsukuba, Japan

From my experiences in research on digital libraries (DLs), I have found the following issues are quite crucial for the development of DLs.

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 make database/IR systems be able to use them for interoperability between legacy and new data.

2. Levels and types of 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, e.g., 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 database 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 for finding information resources. However, on the other hand, quality resources stored in DLs tend to be stored in databases which are not accessible to the robots. This means users searching quality resources may have to search every DL. Therefore, metadata sharing is crucial to enhance accessibility to the quality resources, and collaborative development of the shared metadata by DLs is important.

My current research interest is mainly in metadata, especially in subject gateway and metadata registry. Firstly, subject gateway (SG) is well known as an important function of DLs. An SG provides functions for users to navigate to appropriate information resources in a certain subject area(s). The SG should have not only information about Web resources but also information about digital materials stored in DLs in its subject area. Since SGs have potential for cross-DL (or cross-Archive) searching I think discussion on DLs from the viewpoint of subject gateways is useful for the open archives. Secondly, metadata is a very important component for subject gateways and also for DLs in general. Metadata interoperability and long-term usability are crucial issues. To cope with these issues, metadata registries will play an important role. This is because the schema information stored in the registries is very much useful for users who are engaged in development of DLs and archives. In addition, since every DL (archive) might have its own metadata schema, mapping functions between different schemas, which should be maintained as a shared resource and should be understandable both by humans and machines, will play a crucial role for cross-DL resource access. Metadata registry is also important for keeping track of revisions of metadata definitions, in other words, for chronological interoperability between legacy and new data.

Suleman, Hussein

My interest in Open Archives stems firstly from my past experiences as a student at a South African university. With very limited network resources, online publication databases were and are not very widespread as a research medium. Different techniques like caching and mirroring have been employed in the past to overcome this problem in other domains. Now, the Open Archives initiative offers us a mechanism to easily share databases using protocols that will support mirroring and consolidation of data from remote sites.

Another issue that cannot be ignored if you live in a third-world country is that subscription-based services are relatively much more expensive. Thus, free access to articles or even just the metadata helps enormously in disseminating research results. I am interested in both the technical aspects of making information available as well as the philosophy and economic models on which such systems are based.

At the present moment, I am a member of the Digital Libraries Research Laboratory at Virginia Tech. I am working with various digital libraries that are compliant with the OAi protocols (Web Characterization Repository, Computer Science Teaching Center, NDLTD) and am currently involved in discussions with the administrators of other potentially-compliant archives at Virginia Tech. I have written an alternative Dienst-subset implementation, which is currently used by two of the archives at Virginia Tech, and a test program which comprehensively tests an archive for compliance with the OAI protocols.

I am in the initial stages of work to deal with the problem of management of the massive quantities of data that are made available in an Open Archives environment. The simple solution is to accumulate all metadata from upstream archives, but this suffers from a scalability problem that is an all-too-familiar lesson we have learnt from the World-Wide Web. As the amount of metadata in a digital library grows, it becomes increasingly difficult to find relevant documents through search and browse operations. Imposing structure on the data is one approach to deal with this complexity.

From the perspective of interoperable archives, a specialized archive in a particular subject area could be built by harvesting data from multiple large-scale data repositories such as arXiv. Then each entry in the archive can be filtered to determine relevance in a manner similar to that of the TREC experiments. In a live system, however, it is not always possible to make a permanently binding binary decision about relevance at any one point in time, since as the subject area evolves, so too may the relevance judgements.

With these issues in mind, in the immediate future I plan to study the construction of specialized archives, with metadata harvested from multiple large general-purpose open archives, using a combination of automatic and manual classification techniques to determine a dynamic probabilistic relevance to the subject area that defines the specialized archive. This work will potentially be useful to subject-specific digital libraries that wish to benefit from Open Archives without incurring unnecessary storage costs and without compromising the quality of data in the archives.

Tansley, Robert

I have implemented an Open Archive complying with the Santa Fe agreement for interoperability. The software has been designed to be as flexible as possible while still being easily configurable by other people. The first application of this software is as the replacement of an existing open archive covering the cognitive sciences, CogPrints. My involvement with the OAi is largely technical.

My first concern with the current standard is the vagueness of the partition mechanism. Without a standardised way to relate partitions between archives they are largely pointless. If each archive has its own arbitrary set of partitions, what use are they for interoperability? If I specify a partition during a search I'd only get one partition from one archive. I believe we should consider adopting a logical classification scheme so that a searcher can make meaningful use of partitions.

I have also encountered various ambiguities in the Dienst protocol. For example, the specification does not make it clear whether a partitionspec specifies a single partition (including its ancestry in the partition hierarchy) or more than one. It also does not make it clear whether the scope of a particular partition encompasses other `child' partitions in the hierarchy.

Another issue is that the protocol is inconsistent in using escaped character sequences in requests. Some characters that should be escaped are, and some are not. Behaviour in this area needs to be consistent if open archives are to be able to communicate.

One more issue requiring discussion is I believe how data providers are going to inform service providers of their presence. I don't think putting the details up on a Web page, and having the service providers manually putting details into their software is a scalable approach. Some way of registering a data provider with a "master server" somewhere is one possibility; this could also provide the basis for an archive specifying the scope of its subject matter (a la classification).

Warner, Simeon

(see Ginsparg, Paul)

Wayland, Ross

The University of Virginia Library (http://www.lib.virginia.edu) began its first phase of digital library development by creating the Electronic Text Center in 1992, a pioneering effort to create collections of structured electronic texts, provide them on the Internet and to provide support to faculty and students creating their own collections by providing facilities and training. Soon after, we established centers for digital media, including images, sound and video, for social sciences and geographic information and for special collections, continuing the dual roles of public service and collection building in each center. The Library has earned an international reputation for its work with digital content that has now expanded to include images, sound, video, maps, and numeric data.

Work began in 1998 to develop a vision for building the "Library of Tomorrow", a five-year program to transform the traditional library into the model university research library for the twenty-first century. The Library's second phase began by establishing a team known as the Digital Library Research and Development Group to build a system to manage the burgeoning electronic collections and provide integrated access to them. After investigating existing digital library "solutions" we ascertained that an appropriate system did not yet exist and we developed a conceptual plan for the management system that we know we need. Now we are beginning to work towards assembling a system by using existing tools where we can find them, developing them where we cannot.

We are interested in following emerging technologies, protocols, and standards involved in the indexing, storage, retrieval, and dissemination of digital information.

I will be representing the Library's Digital Library Research and Development Group at the OAI meeting on Saturday.

Ross Wayland
Email: rlw@virginia.edu University of Virginia

Phone: 804-924-0746 Alderman Library – Systems
FAX: 804-924-1431 Charlottesville, VA 22903-2498

Technical Issues for San Antonio Open Archives Meeting

Carl Lagoze

A primary goal of the San Antonio Open Archives meeting should be a thorough review and refinement of the technical agreements that comprise the Santa Fe Convention. These agreements were reached somewhat hastily at the Santa Fe meeting and experience has shown that there are a number of problems with them. Since adoption of the Convention is still rather limited, readjustments would not present an undue burden. This will not be true in the future with (hopefully) wider adoption.

The remainder of this document describes areas where improvements are needed, based on our experience at Cornell.

Unique Identifiers

The convention mandates a *record identifier* consisting of a prefix *archive identifier* and a *local identifier*. The archive identifier must be all alpha-numeric and the first non-alpha-numeric character is then used as the prefix to suffix separator. There are three problems with this:

1. The relationship between Open Archives identifiers and URI's, as defined by RFC 2396, is problematic. For example, the DOI 10.1045/may2000-kaser, while syntactically a correct Open Archives identifier, has a problem since the DOI prefix (10.1045) is not recognized as an OAI prefix (OAi providers will consider "10" the prefix). It seems inappropriate to effectively penalize those archives that use real URN's in the effort to create OAi URNs.
2. Related to the above, those of us who do use real URNs (such as NCSTRL which uses handles for its naming scheme) have archives with records that have multiple prefixes. For example, in the Cornell NCSTRL repository (at cs-tr.cs.cornell.edu) we have records with handles that are prefixed by both ncstrl.cornell and ncstrl.cornelltc. Again, the OAi identifier scheme penalizes archives like the NCSTRL members (or D-Lib Magazine).
3. Again related to the above, those of us who do use real URNs have prefixes that appear across multiple archives (repositories). In fact, we have documents with the same URN that are mirrored across multiple archives (this is definitely true of arXiv).

I believe we need to solve these problems. The goal of unique identifiers for records makes sense. It doesn't make sense to do it in a fashion that penalizes those archives that actually implement unique identifiers. Here are some suggestions for fixes to the problem, none of which solve the entire problem. First it seems that we should adopt '/' as the separator character between the prefix and the suffix. Second, we should allow archives to register (or request as shown next) multiple prefixes for an archive. Finally, we might consider assigning prefixes rather than allowing requests. The latter doesn't really scale since inevitably we will end up with collisions, while the former guarantees no collisions (since we assign them). We might even consider getting assignments from the DOI organizers or from the CNRI handle authorities.

Metadata

There is no perfect or correct metadata set. I don't think that we should spend our time in San Antonio arguing endlessly about the composition (semantics) of the basic metadata set. However, I do think that we made an error in not using the Dublin Core Metadata Element set (DCMES) where possible. Simply adopting the DCMES is not wise, since it is both too undefined (all elements optional, all elements repeatable) and too extensible (the notion of unlimited qualifiers). We should, however, adopt an application profile (as my colleague Tom Baker calls it) that is a well-constrained subset of the DCMES mixed with elements, when needed, from another namespace. The table below shows mapping of OAMS metadata elements to the DC elements (some with qualifiers from the "basic qualifier set" for the Dublin Core defined at http://www.mailbase.ac.uk/lists/dc-general/2000-04/0010.html). Blanks in the "DC Namespace" column indicate where there is no good mapping. The one slightly problematic mapping is

“Author” to “Creator”, because we have decided to include organizational affiliation with each author. This could be solved by creating in the XML an OAMS:Author elements with two interior elements; DC:Creator and OAMS:Affiliation.

OAMS Namespace	DC Namespace
Title	Title
Accession	Date (qualified with “available”)
Display ID	
FullID	Identifier
Author	Creator
Abstract	Description (qualified with “abstract”)
Subject	Subject
Comment	
Discovery	Date

Protocol

The protocol is lacking two functional components:

1. The protocol does not include the ability to indicate removed records. The List-Contents verb should include notation to indicate that a record has been removed.
2. Many sites will want to return a limited set of records in the List-Contents response. The response should be able to indicate “ask for more with this argument” so that clients will know that all records have not been returned and that they can get additional records with the corresponding argument.

Review of OAi NASA Langley Research Center / Old Dominion University's Implementation Experiences and Directions

Michael L. Nelson <m.l.nelson@larc.nasa.gov>
Kurt Maly <maly@cs.odu.edu>
Mohammad Zubair <zubair@cs.odu.edu>

Our experiences in building the Universal Preprint Service (UPS; http://ups.cs.odu.edu/), the prototype demonstrated at the initial Santa Fe meeting, have led us to pursue two research tracks related to the Open Archives initiative (OAi).

Implementation of the Dienst Protocol Using Oracle

The UPS prototype uses NCSTRL+, a slightly modified version of the Dienst protocol to build a cross-archive digital library (DL). UPS has nearly 200,000 records harvested from six production archives. We ran into a number of problems while using the current Dienst software in building a DL of that size:

- A system limit in Solaris prevents a directory from having more than 32,767 inodes (files or subdirectories). This limited the current version of Dienst to 32,767 records per publishing authority.
- The freeWAIS-sf search engine used by the current Dienst implementation has two significant limitations: 1) the number of hits is limited to approximately 250; 2) if a word occurs more than 20,000 times, it is considered a stopword, regardless of the size of the collection.

These limitations proved to be serious in constructing a large-scale DL. To address these issues, we have begun implementation of the Dienst protocol using a combination of Java search servlets that search an Oracle RDBMS. Additional optimizations include a session manager and caching of searches for increased performance. A technical report has been issued to cover the details of this implementation.

Implementation of the OAi Dienst subset protocol using Buckets

Buckets are aggregative, intelligent digital objects for publishing in digital libraries. Each of the records in UPS is stored in a bucket. The UPS buckets are called "lightweight buckets", because their content still resides in the original archives. Buckets proved to be convenient mount points for value added services, such as the SFX reference linking service. Because the buckets are modifiable, we are designing "dumb archive" (DA) buckets (from the "Smart Object, Dumb Archive" (SODA) model for DLs) that will map the Open Archives Dienst Subset archive messages into the existing DA methods for manipulating archives. This will provide an alternate implementation of the OAi harvesting protocol.

Please see our website, http://dlib.cs.odu.edu/, for the latest updates on our results.

Report on Open Archives Work at Virginia Tech

Hussein Suleman (hussein@vt.edu)

Current OAi-Compliant Archives at Virginia Tech

1. Computer Science Teaching Center (CSTC)
   Digital library of peer-reviewed teaching resources for computer science educators and learners.
   http://www.cstc.org/
2. Web Characterization Repository (WCR)
   Online database of metadata for resources (traces files, publications, tools) in the field of Web Characterization.
   http://purl.org/net/repository/
3. Virginia Tech Electronic Thesis and Dissertation Collection (VTETD)
   Collection of theses and dissertations produced and stored online at Virginia Tech.
   http://scholar.lib.vt.edu/theses/

Related Work

1. Repository Explorer
   An interactive compliancy test for Open Archives, aimed at technical personnel involved with implementation.
   http://purl.org/net/explorer/
2. USMARC XML-DTD
   Development of an XML DTD for the MARC standard to be used by the OAi.
   http://www.dlib.vt.edu/projects/OpenArchives/index.html
3. Alternative Perl implementation of the Dienst subset protocol.

Work in Progress

1. NDLTD (Networked Digital Library of Theses and Dissertations)
   Integration of various interoperable thesis/dissertation archives, including OA-compliant ones, into Marian, an information retrieval / digital library system developed at VT.
   http://www.ndltd.org/
2. Alternative Java implementation of the Dienst subset protocol.

Prospective Projects

1. Making VT mirror of Project Gutenberg into an OAi-compliant archive.
2. Making our library collection of locally-hosted electronic journals accessible through OAi protocols.

Discussion Topics

Organizational Issues

1. When encountering technical problems, we found that in many instances the organizational structure of the OAi did not have the support mechanisms to deal with the technical problems. There were no individuals or committees who had ownership of the protocols and specifications - thus no specific person could clarify ambiguities and many of these are to this day unresolved. It would greatly help implementers if particular aspects of the technical specifications are moderated by definite people so that problems can be resolved in the short term.

2. It has also been noted that there do not exist forums for public discussion separate from the UPS listserv. This is fairly unusual to implementers familiar with the Open Source movement, who are used to being part of an open-membership peer group that attempts to solve problems as they occur.
3. Another specific problem with the protocol documentation has to do with versioning. While some version information does exist, the versions of various documents still do not correspond. For example, the XML-DTD (which I would expect to be the authoritative document for the metadata standard) is in fact not in sync with the current Dienst subset documentation.
4. In general, the feeling is that a better organizational structure will go a long way to provide better support for the resolution of current and future technical issues.

Technical Issues - Concerns and Unresolved Problems

1. Partition Specifications

1.1. Does a partition specification list all interior nodes or just the leaf?

1.2. If all interior nodes are listed, should there not be two separators - one for nodes within a path and one for paths within a list?

1.3. Do we allow overlapping partitions? (Some of our archives allow this while others do not.)

1.4. Do we allow multiple partitionings of the same space (CSTC currently does this)? If so, is the intention (implicit or explicit) to have partitions correspond to subject areas for objects?

1.5. To further automate the process, how about a verb to return what the separator/s is/are? (possibly as a Unicode entity)

2. Updates and Deletions

2.1. We assumed that updated information will get resent automatically since the date will be updated and when a service provider receives the same ID multiple times from a single source it can consider all data after the first set to be updates. What happens if the date is not updated?

2.2. How do we handle deletions from a database?

3. Namespaces

3.1. There are various places in the specifications where references are made to "oams" and "OAMS" - which of these is the proper name for the metadata set? (Note that in this case the Dienst spec does not agree with the acceptable metadata listing.)

3.2. Do we prepend the metadata set name to each element? (The Dienst protocol does it but the XML-DTD doesn’t.)

3.3. How do we ensure ownership of objects when the metadata is exported to a remote database? Does the full-id never change? What if two archives export the same object's metadata with different identifiers to a third database - what does the third database subsequently export?

4. Versioning

4.1. Can we have version numbers related to the whole protocol rather than individual verbs? (so as to reduce complexity)

4.2. Can we version the metadata internally (within the XML) so that changes do not disrupt existing systems?

5. Listings

5.1. Can we produce sectional listings that end on a specific date? a "to-date"?

5.2. How about a verb to return a count of the (estimated/actual) number of records?

5.3. How can we guarantee completeness of harvesting without locking the source database? If we store the date before asking for an update then we may get overlapped records - if we store the date afterwards, then we may miss records.

6. Character Sets / Markup

6.1. Do we support standard SGML character sets in the generated XML? There are individual sets proposed for different DTDs - can these be generalized? What happens when an archive (like Marian) wants to export records with differing formats?

6.2. Do we adopt the use of ISO8879 and/or Unicode entities? What are the implications for parsers and the level of complexity we require from any service provider? (For NDLTD we have converted all entities to Unicode because of this concern.)

6.3. Do we allow HTML markup in abstracts for the purpose of paragraphing? What about other formatting? If we strip the HTML codes, is there an independent syntax to format abstracts or do we expect a single paragraph? Alternatively, do we stick to simple CRLF sequences?

7. Special collections

7.1. The collections at Virginia Tech are not primarily archives of articles. WCR contains data sets, CSTC contains software/courseware, and VTETD contains theses and dissertations. What are the implications of different types of collections for the OAi? Should the specifications be generalized to deal with such things (for example, an abstract has no meaning for a web trace file that is being exported)? Is Dublin Core a better choice of metadata set because of its generality?

7.2. Should we allow each archive to export unique metadata sets? All VT archives have additional metadata that is lost through the export process. Should we have basic sets with optional name/value pairs? Is it possible to establish a semantic-level metadata mapping database (even for specific cases)?

7.3. Do we allow for bibliographies? One very popular website at VT is just a listing of references in the field of feminism - should we consider making this compliant, and what are the implications?

8. Dienst subset

8.1. Why are fixed parameters used instead of named ones? This is not orthogonal across the protocol and does not conform with common CGI practice - it may be problematic since many system that process URLs (proxy servers, log analyzers, etc.) assume parameters are only those found after a "?".

8.2. Some URLs contains entities but others don't. Can this be standardized?

8.3. Instead of expecting people to enter information about the archive into a web page, how about a verb that returns the information? Then the URL can be sent to the Open Archives website, which can call the verb to display information. This automation will help systems that want to import the data.

9. Data Cleaning

9.1. How much cleaning should be done before exporting data? We don't want ripple effects to further decrease the quality of data as character sets are changed and other modifications are made. We perform a substantial amount of cleaning on the VTETD collection before exporting the data.

9.2. How do we handle whitespace? Some existing archives use lots of unnecessary whitespace in the XML and this has to be carefully removed when parsing - a cleaner XML file will make it easier to import the data.

Santa Fe Convention : Core Document

This core document presents a step by step approach for making your e-print archive or your service comply with the Santa Fe Convention. To clarify the Santa Fe guidelines, some underlying concepts are introduced first. Technical details are in separate documents, to which links are provided within the text and in a list at the end of this document.

Outline

- ♦ Introductory concepts
  - Archives and open e-print archives
  - Managed e-print archives
  - Data providers and service providers
  - Data in an e-print archive
- ♦ For the data provider: how to make your e-print archive comply with the Santa Fe Convention?
  - Step 1: Choose a unique identifier for your e-print archive
  - Step 2: Use unique persistent identifiers for data in the archive
  - Step 3: Implement the Open Archives Metadata Set
  - Step 4: Implement and document other metadata formats supported by your e-print archive
  - Step 5: Implement the Dienst harvesting interface
  - Step 6: Let the Open Archives initiative know that your e-print archive is open
- ♦ For the service provider: how to make your services comply with the Santa Fe Convention?
  - Step 1: Retain the original identifiers in your services
  - Step 2: Comply with the usage restrictions specified by the data providers
  - Step 3: Let the Open Archives initiative know that you have developed a service based on open archives data

Introductory concepts

Archives and open e-print archives

We consider the following to be crucial components of an e-print archive:

- A submission mechanism
- A long-term storage system.

In addition, we consider it crucial that an e-print archive be open, incorporating a mechanism that enables third parties to collect data from the archive. Such a mechanism allows third parties to create end-user services that support the discovery, presentation and analysis of data in the archive. We recognize that most e-print archives will also provide end-user services. However, we consider that facilitating the broad dissemination of archive data through third party services is a crucial feature of an e-print archive.

Managed e-print archives

We also assume that e-print archives are managed. This means that they have some form of policy with regard to the submission of documents and also a policy with regard to the preservation and retention of documents.

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, we make a distinction between data providers and service providers.

- 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

An archive may store metadata that describes full content without storing the full content itself. In this case, we consider the metadata as a record. However, we assume 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 we consider the combination of metadata and full content as a *record*.

In this convention, therefore, we define an archive as a collection of records. These records have the following properties:

- A record in an e-print archive contains, at least, metadata that describes full content
- A record in an e-print archive may also contain full content such as a research paper, a dataset, software, etc. or a bundle of these.

For the data provider: how to make your e-print archive comply with the Santa Fe Convention?

Step 1: Choose a unique identifier for your e-print archive

To support interoperability, each archive should have a *unique archive identifier*. This identifier refers to the authority managing the archive or to the archive initiative. Choose an identifier that consists of alphanumerical characters [a-z, A-Z, 0-9]. To make sure that your archive identifier does not coincide with that of another archive, check the list of existing identifiers at http://www.openarchives.org/sfc/sfc_archives.htm. Formally, the case of characters in the archive identifier is significant however identifiers should be selected to be distinct regardless of case.

- *When setting up an open archive for the University of Spa in Belgium, one could choose BESPA as the archive identifier.*
- *The existing NCSTRL initiative could choose NCSTRL, while the RePEc initiative could opt for RePEc.*

Step 2: Use unique persistent identifiers for records in the archive

Records in your archive should have unique persistent identifiers. It is up to you to make sound decisions on their structure and generation.

When you combine a *unique archive identifier* and a *unique record identifier* for a record in your archive, the result is a *full identifier* for a record in your archive, that will never coincide with a full identifier of a record in another archive. This makes the job of service providers a lot easier. Choose a printable, non-alphanumeric character as a *separator* to delimit the archive identifier from the record identifier.

- *The archive of the University of Spa will accord meaningful identifiers to metadata submitted to the archive. Let us assume that those identifiers consist of the faculty of the first author followed by a sequence number that contains date information. Hence, a record identifier might be MEDICINE/19991104/012. If this archive chooses a hyphen to delimit the archive identifier from the record identifier, the full identifier for the record would be: BESPA-MEDICINE/19991104/012.*
- *arXiv.org uses record identifiers that start with the name of a sub-archive followed by a date-sensitive sequence number. For instance, physics/9811004 and hep-th/9909044 are valid identifiers in this archive. In this case, using a colon as delimiter, the full identifiers would become arXiv:physics/9811004 and arXiv:hep-th/9909044.*

The identifier of a record in your archive - either the full identifier or the record identifier (without its leading archive identifier) - will be the crucial key for extracting the metadata for a record. In some archives, it may also be the key to get to the full content of the record. In other archives, other metadata elements within a record will point to the full content.

Step 3: Implement the Open Archives Metadata Set

We recognize that archives will use specific metadata sets and formats that suit the needs of their communities and the types of data they handle. However, interoperability depends on a shared format for exchanging metadata and therefore archives should implement the basic Open Archives Metadata Set (oams) which is described at http://www.openarchives.org/sfc/sfc_oams.htm.

Step 4: Implement and document other metadata formats supported by your e-print archive

Service providers will be able to provide more powerful services for users if a metadata format that is richer than the basic oams can be harvested from your archive. We encourage data providers to provide access to the full richness of metadata available to support discovery and retrieval of records in their archive, preferably by adopting an exchange format already used by other e-print archives. To help you determine whether an existing format can serve your needs, we maintain a list of such metadata formats at http://www.openarchives.org/sfc/sfc_metadata.htm.

If no existing format suits your needs, you must take on the tough task of compiling your own format. For compliance with the harvesting interface presented in Step 5, you will also have to define an XML representation for it. It is important that you document your metadata format fully and make the documentation available to service providers. Again, it will make their jobs a lot easier. If you inform the Open Archives initiative about the alternative metadata format for your archive, we will include it on the OA list of metadata formats at http://www.openarchives.org/sfc/sfc_metadata.htm, so that others can benefit from your work.

- *The formats that are in use or under implementation by the archives contributing to the compilation of this convention are - at the time of writing - the following: MARC, ReDIF, Dublin Core, REFER, RFC 1807, Open Archives Metadata Set.*

Step 5: Implement the Dienst harvesting interface

Once your archive has identifiers and supports one or more metadata formats, the next step turns it into an open archive, by making sure that service providers can access data from your archive. Because we anticipate the creation of many open archives and because we want services across those archives to be built in a short to medium time period, we strongly recommend that all archives implement the same harvesting interface. Such a harvesting interface must allow third parties to write software that collects data selectively from the open archives. We propose a harvesting interface that complies with the Open Archives Subset of the Dienst Protocol, which is described in detail at http://www.openarchives.org/sfc/sfc_dienst.htm. To support a better understanding of the recommendations in this convention, we provide a brief description of the protocol subset here.

The Dienst protocol has an HTTP-based implementation. Its Open Archives subset defines a communication procedure, as well as the syntax for the corresponding messages and responses, that will allow service providers to harvest metadata selectively from open archives that comply with the Santa Fe Convention. The procedure has three steps:

- **Action 1: an archive can be polled to obtain the following information about the archive:**
  - The logical partitions into which records are grouped within the archive
  - The metadata formats that are supported for delivery of archive metadata in response to a harvesting request.
- The Open Archives Subset of the Dienst protocol defines how such polling requests should be sent and the syntax used by an archive to respond to such requests. It does not define the set of valid responses to the metadata format request. But in the Open Archives context, the list of valid formats and their identifiers is available at http://www.openarchives.org/sfc/sfc_metadata.htm.
- Archives should support several criteria to divide their content into logical partitions that are recognized by the Open Archives Dienst Subset. We especially recommend subject-oriented partitioning as well as partitioning based on author affiliation.
- **Action 2: a list of identifiers for records in an archive can be requested.**
- We expect these identifiers to be the unique persistent identifiers of the records in the archives. The Open Archives Dienst Subset defines the syntax to request:
  - A list of identifiers for all records in the archive
  - A list of identifiers for the records in a partition of the archive
  - A list of identifiers for records that have become available in the archive after a specified date
  - A list of identifiers for records that have become available in an archive partition after a specified date.
- The Open Archives Dienst Subset also defines the way in which the list of identifiers will be returned.
- **Action 3: given a list of identifiers (obtained by Action 2) and a supported metadata format (identified through Action 1), a request to harvest metadata can be sent.**
- The Open Archives Dienst Subset defines the syntax for the harvesting request. In the subset, exchange of metadata is always in XML, whichever metadata set is chosen. A list of formats and details on their interpretation and XML representation is available at http://www.openarchives.org/sfc/sfc_metadata.htm. An archive should respond to a request for metadata in a specified format by returning data rendered according to the corresponding exchange format.

Step 6: Let the Open Archives initiative know that your e-print archive is open

The last step is to join the community of open e-print archives by informing the Open Archives initiative of the details for your open archive. You should use the template we provide for that purpose at http://www.openarchives.org/sfc/data_provider_template.htm to do so. Fill out the appropriate information, install it on your archive server and send us an e-mail at openarchives@openarchives.org to inform us of its URL. We will include your archive in the list of Santa Fe compliant archives that we maintain at http://www.openarchives.org/sfc/sfc_archives.htm, thus helping to make your archive conveniently visible to service providers. You will see that the form also includes a possibility to inform service providers of restrictions that apply in the usage of your data.

You may also want to insert the Open Archives logo in the main entry point of your archive to indicate that you have joined the initiative.

For the service provider: how to make your services comply with the Santa Fe Convention?

As you can see from the recommendations we make to data providers, our aim is to make it easy for you - the service provider - to create services based on open archive data. In return, we ask you to comply with the following:

Step 1: Retain the original identifiers in your services

When you create services based on records originating from open archives, keep the original full identifiers associated with the data as a means of indicating the provenance of records.

Step 2: Comply with the usage restrictions specified by the data providers

With regard to the usage of data from the open archives, respect the restrictions that data providers mention in the form describing their archive. For each open archive, this form is accessible via http://www.openarchives.org/sfc/sfc_archives.htm.

Step 3: Let the Open Archives initiative know that you have developed a service based on open archives data

Inform the Open Archives initiative and thus the data providers about the data that you are harvesting and about the use you make of it. To facilitate this, we provide a template at http://www.openarchives.org/sfc/service_provider_template.htm that you are encouraged to use and fill out. Install it on your server and inform us of its URL by sending us e-mail at openarchives@openarchives.org. By doing so, you are joining the community of Open Archives. We will list all information on service providers at http://www.openarchives.org/sfc/sfc_services.htm.

You may also want to insert the Open Archives logo in the main entry point of your service to indicate that you have joined the initiative.

January 19th 2000

The Open Archives Dienst Subset

Document Version 2000-05-19 08:19:56 -0400

Jim Davis (jrd3@alum.mit.edu)
David Fielding (fielding@cs.cornell.edu)
Carl Lagoze (lagoze@cs.cornell.edu)
Richard Marisa (rjm2@cornell.edu)

0. Changes

Any changes to this specification after February 15, 2000 will be noted here.

- Updated `oams` meta format in examples. This document was written before the final `oams` meta format was adopted so the examples that return `oams` meta data were incorrect.
- Structure version number was incorrect. Should be 2.0 instead of 1.0.
- Structure verb may return multiple meta data formats. Changed <meta-format> tag to <meta_formats>.

1. Introduction

This document describes the portion of the Dienst protocol that is used for basic interoperability within archives in the Open Archives initiative, as recommended in its Santa Fe Convention. The goal of the Open Archives initiative is to provide the mechanisms for interoperability among distributed e-print archives. The protocol described in this document allows harvesting of metadata for uniquely identified *records* in an archive. The word *document* is purposely avoided and the notion of a *record* is purposely imprecise. Some archives may just provide access to metadata, others may also provide access to metadata and full content in some form, others may provide other services associated with the metadata and content such as access to the full content in various manifestations (formats) or structural decompositions (e.g., individual pages, chapters, and the like).

The protocol described in this document is a subset of the full Dienst protocol, which provides for communications with services in a distributed digital library. When this subset of Dienst needs to be differentiated from the full Dienst protocol, it will be referred to as the *Open Archives Dienst Subset* for the remainder of this document. Readers will notice the use of the word *Repository* in the Dienst protocol requests. This follows from the use of the term *Repository* in the broader Dienst system in lieu of the term *Archive*.

2. Protocol Features

2.1 Unique Identifiers

All archives participating in the Open Archives Initiative have a *unique archive identifier*. This identifier is restricted to alphanumeric characters. Registration of this identifier is part of the Open Archives registration process for data providers, described in Step 6 of the Santa Fe Convention of the Open Archives Initiative. All records in an archive have a *unique record identifier* - unique within the scope of that archive. These two identifiers - the unique archive identifier and the unique record identifier - can then be concatenated (separated by any printable non-alphanumeric character) to form a unique *full identifier* (referred to as a `fullID` in the protocol documentation). For example, the unique archive identifier `handlecorp` can be combined with the unique record identifier `11223` and separated with the `/` (slash) character to form the full identifier `handlecorp/11223`. The full identifier is then used and returned by Dienst requests.

2.2 Partitions

The Dienst protocol defines the notion of a **`partition`** within an archive**.** A partition is an administrator-defined subset of the archive. Each partition has a (one token) `name` and a (possibly) longer `description`. Depending on the policy of an archive an individual record may exist in one or more partitions. Note that there is, in general, no way to predict the partition in which a record appears from its full identifier, or even given full knowledge of the record.

An archive may have one or more partition hierarchies. For example, an administrator may decide to partition an archive into two hierarchies, one based on institutional affiliation and one based on subjects as follows:

- Institutions
    - Oceanside University of Nebraska
        - Department of Computational Entomology
        - Department of Metaphysical Phenomenology
    - Valley View University of Florida
        - Department of Frenetics
        - Department of Histrionics
- Subjects
    - Existential Kenesiology
    - Quantum Psychology

The partition hierarchies in an archive are available via the `List-Partitions` request .

2.2.1 Partition specifications

The `List-Contents` verb includes, as an argument, a `partition specification`. Partition specifications are expressed in the following grammar where `partitionname` is the short one token name for the partition:

```
partitionspec := partitionlist
partitionlist := partitionsel | partitionsel;partitionlist
partitionsel  := partitionname
partitionname := [A-Za-z0-9-_]+
```

Example:

```
Institutions;Florida;Frenetics
```

Where Florida is the short name for the partition Valley View University of Florida and Frenetics is the short name for the partition Department of Frenetics.

2.3 Verbs and Versions

Individual Dienst protocol requests are called *Verbs*. There may be more than one version of a verb, with each version differing in syntax or semantics. A version takes the form of two integers, separated by a period. This version applies to the individual verb, not the protocol as a whole. (The protocol as a whole does not have a version number. The date on the protocol document indicates the set of verbs that are defined as of that date.) Including a version number in the message allows for backward-compatible extension to the Dienst system.

An archive might support verbs in various versions. An archive receiving a message with an older version number must either reply using the old syntax and semantics, or reply with an error. If an archive receives a message with a *newer* version number, then it must return an error.

Software supporting the Open Archives Dienst Subset may or may not be versioned. If a software version number exists, that number is independent of the Dienst protocol verbs and versions of those verbs that the software supports.

2.4 HTTP embedding of Dienst requests

Dienst protocol requests are expressed as URLs embedded in HTTP requests. A typical implementation uses a standard Web server, such as Apache, that is configured to dispatch Dienst URLs to the software implementing these requests. The remainder of this section describes the aspects of the protocol that are specific to the HTTP embedding.

2.4.1 Message format

All messages are encoded into URLs where the path portion of the URL consists of the following tokens, in the following order:

Dienst

> This token appears literally in the URL.

Service Name

> The name of the service which is to handle the message. The only service implemented in Open Archives Dienst Subset is Repository.

Version

> The version of the verb being invoked.

Verb

> This is the name of the message, e.g. List-Contents. A verb is unique within a Service.

Fixed arguments

> Each verb has a certain number of fixed arguments, which must always be supplied, and must appear in the order cited.

Keyword arguments

> Keyword arguments take the form key=value. If there is more than one keyword argument, they are separated by an ampersand. Arguments may appear in any order. Unless specified,

keyword arguments are always *optional.*

The separator between tokens in the path is the slash, except that the separator before the keyword arguments is a question mark.

Example

If the `Repository` service implemented the `Shred` verb, and if version `1.2` of that verb accepted two keyword arguments (`delay` and `volume`), then an example request is:

```
/Dienst/Repository/1.2/Shred?delay=9&volume=7.4.
```

The full URL for this request at a particular Web server might be:

```
http://bar.com/Dienst/Repository/1.2/Shred?delay=9&volume=7.4.
```

2.4.2 Special characters

The syntax rules for URIs restrict a few characters to special roles in certain contexts and require that if these characters are used in any other way that they be written as an escape sequence; a percent sign followed by the character code in hexadecimal. The reserved characters are.

Character	*Role*	*Escape Sequence*
`/`	Path Component Separator	`%2F`
`?`	Query Component Separator	`%3F`
`#`	Fragment Identifier	`%23`
`=`	Name/Value Separator	`%3D`
`&`	Argument Separator in Query Component	`%26`
`:`	Host Port Separator	`%3A`
`;`	Authority Namespace Separator	`%3B`

Finally, the space character may not appear anyplace in a URL. It must be written with a "`+`" (or with the percent sign escape sequence `%20`.)

As a result, use of these characters **must** be escaped within a Dienst protocol request if their use does not correspond to their established URI role. Note that in the examples used throughout this document, special character escaping is shown.

2.4.3 Message Responses

Responses to messages are formatted as HTTP responses, with appropriate HTTP header fields. The *`return type`* specified for each protocol request in this document will, therefore, correspond to the MIME type included in the HTTP `Content-Type` header field

2.4.3.1 MIME Types

The responses to all Open Archives Dienst Subset requests are structured streams with MIME type `text/xml`. An appendix to this document lists the DTD (Document Type Definition) for every verb. All XML responses to Dienst protocol requests have the following uniform features.

I. The first tag output is a XML declaration where the `version` is always `1.0` and the `encoding` is always `UTF-8`.

II. The remaining content is enclosed in a root element that has the same name as the verb of the respective request. The element has a single attribute named `version`, which has a value that is the version of the verb of the respective request. For example, a `Disseminate` verb with version `2.0` will produce `text/xml` content with an wrapped in a tag like
`<Disseminate version="2.0">`

2.4.3.2 Status Codes

Status codes and error returns correspond to those defined for HTTP (refer to that protocol documentation). A normal response from a Dienst message in HTTP is signaled with the `200` reply code. Error returns are signaled with the appropriate `4xx` code as specified in the HTTP protocol. The use of HTTP error codes is as follows:

- `400` - if the Dienst request is malformed; for example, illegal arguments or the values of arguments are invalid.
- `404` - if a record specified in a Dienst request is not in the archive.

For each error return, the HTTP `reason-phrase` returned with the code should provide additional information useful to a human reader.

2.5 Dates

All dates in the protocol (requests and responses) are encoded using the "Complete date" variant of ISO8601. This format is `CCYY-MM-DD` where `CC` is the century, `YY` is the year, `MM` is the month of the year between 01 (January) and 12 (December), and `DD` is the day of the month between 01 and 28 or 29 or 30 or 31, depending on length of month and whether it is a leap year.

3. Protocol Messages

This section lists the messages (verbs) implemented by the Open Archives Dienst Subset. Each message has a `Name` (which is used for purposes of discussion), a `Verb` (a unique name for the message, used in the protocol to name the message), a `Version`, a list of `Fixed arguments`, a list of `Keyword arguments`, a `Return MIME type` and return `status codes`. The documentation for every message includes an example request and response (where appropriate) and the meaning of HTTP error codes that may be returned. These examples uniformly use the full identifier `handlecorp/970101`.

To make reading of this document easier, the DTDs for responses to verbs that return `text/xml` are separated from the main body of the document into an appendix.

Disseminate Metadata for a Record

Verb: **`Disseminate`**
Version: 1.0
Fixed args: `fullID, meta-format, content-type`
Keyword args: none
Return MIME type: `text/xml`
Return Status Codes: 200, 400, 404

Request the metadata in a specific format from a record.

In addition to the `fullID` , the required fixed arguments are:

- `meta-format` specifies the type of metadata requested. This argument must take the form `#<meta-format>` ,where `<meta-format>` is one of the supported metadata formats (as returned from the `List-Meta-Formats` request). For example, `#oams` requests the Open Archives Metadata Set metadata record for the record instance. The metadata formats supported by an archive can be retrieved using the `List-Meta-Formats` request. The metadata formats available for a particular record can be retrieved using the `Structure` request.

- `content-type` specifies the MIME-type of the content in the dissemination. At this time the only `content-type` that is supported is `xml`.

Example Request:

```
Dienst/Repository/1.0/Disseminate/handlecorp/970101/%23oams/xml
```

Example Response:

```
<?xml version="1.0" encoding="UTF-8"?>
  <Disseminate version="1.0">
    <oams:oams xmlns:oams="http://www.openarchives.org/sfc/sfc_oams.htm">
      <oams:title>A protocol for Interoperable Archives</oams:title>
      <oams:accession date="1994-06-24" />
      <oams:fullId>ncstrl.cornell/TR94-1418</oams:fullId>
      <oams:author>
        <oams:name>James R. Davis</oams:name>
        <oams:organization>Xerox</oams:organization>
      </oams:author>
      <oams:author>
        <oams:name>Carl Lagoze</oams:name>
        <oams:organization>Cornell</oams:organization>
      </oams:author>
    </oams:oams>
  </Disseminate>
```

List Contents

Verb: **`List-Contents`**
Version: 4.0
Fixed args: none
Keyword args: `partitionspec`, `file-after`, `meta-format`
Return MIME type: `text/xml`
Return Status Codes: 200, 400

Return a structured list of the full identifiers for records stored in this archive. Without any arguments the list includes all stored records.

The meaning of the keyword arguments is as follows:

- `file-after` is an optional argument that limits the list to those full identifiers for records that were added or modified since `date`, a universal date expressed in ISO 8601 format. If the server is not able to determine date of modification to the resolution of a day, or if the server is not able to selectively extract records on a time scale of a day, the server may return additional identifiers, e.g. all those modified during the week, month, or even century containing the date.

- `partitionspec` is an optional argument that limits the returned full identifiers to those in the specified partition specification. The partitions available for an archive are returned from the `List-Partitions` request.

- `meta-format` is an optional argument which specifies that, in addition to the full identifier, metadata for each record should be returned in the specified format. The metadata is empty for any record that does not have metadata in that format. The metadata formats supported by an archive can be retrieved using the `List-Meta-Formats` request. The metadata formats available for a particular record can be retrieved using the `Structure` request.

Example Request:

List the full identifiers of records added or modified after January 15, 1998 in the high energy (hep) partition within the physics partition.

```
/Dienst/Repository/4.0/List-Contents
       ?partitionspec=physics;hep&file-after=1998-01-15
```

Example Response:

```
<?xml version="1.0" encoding="UTF-8"?>
  <List-Contents version="4.0">
    <record>arXiv:hep-th/9801001</record>
    <record>arXiv:hep-th/9801002</record>
  </List-Contents>
```

Example Request:

List the Open Archive Metadata Set format along with the full identifiers

```
/Dienst/Repository/4.0/List-Contents
        ?partitionspec=physics;hep&meta-format=oams&file-after=1998-01-15
```

Example Response:

Note that every record includes an `oams` metadata record. If another meta-format were requested (e.g., rfc1807) there might be instances where an empty metadata record was returned (with no data between the metadata format tags) indicating that there is no metadata in that format for the record.

```
<?xml version="1.0" encoding="UTF-8"?>
  <List-Contents version="4.0">
    <record>
      ncstrl.cornell/TR94-1418
      <oams:oams xmlns:oams="http://www.openarchives.org/sfc/sfc_oams.htm">
        <oams:title>A protocol for Interoperable Archives</oams:title>
        <oams:accession date="1994-06-24" />
        <oams:fulId>ncstrl.cornell/TR94-1418<oams:fullId>
        <oams:author>
          <oams:name>James R. Davis</oams:name>
          <oams:organization>Xerox</oams:organization>
        </oams:author>
        <oams:author>
          <oams:name>Carl Lagoze</oams:name>
          <oams:organization>Cornell</oams:organization>
        </oams:author>
      </oams:oams>
    </record>
    <record>
      hdl://cnri.dlib/june96-varian
      <oams:oams xmlns:oams="http://www.openarchives.org/sfc/sfc_oams.htm">
        <oams:title>Pricing Electronic Journals</oams:title>
        <oams:accession date="1996-06-24" />
        <oams:fullId>hdl://cnri.dlib/june96-varian<oams:fullId>
        <oams:author>
          <oams:name>Hal R. Varian</oams:name>
          <oams:organization>UC Berkeley</oams:organization>
        </oams:author>
      </oams:oams>
    </record>
  </List-Contents>
```

Get Metadata Formats

Verb: **`List-Meta-Formats`**

Version: 1.0

Fixed args: none

Keyword args: none

Return MIME type: `text/xml`

Return Status Codes: 200, 400

Returns the metadata formats that are supported by this archive. Note that the fact that a metadata format is supported does *not* mean that it is available for all records in that archive. For each metadata format, the following information is returned:

- The `name`, which is the identifier for the metadata format that can be used in other Dienst protocol requests. A list of identifiers of metadata formats that are in use in the Open Archives context is provided at the Open Archives Web Site.
- The `namespace ID`, which is a URL that refers to a document describing the metadata format.

Example Request:

```
/Dienst/Repository/1.0/List-Meta-Formats
```

Example Response:

```
<?xml version="1.0" encoding="UTF-8"?>
  <List-Meta-Formats version="1.0">
    <meta-format name="rfc1807"
       namespace="http://info.internet.isi.edu/in-notes/rfc/files/rfc1807.txt"
/>
    <meta-format name="dc"
       namespace="http://purl.org/dc" />
    <meta-format name="oams"
       namespace="http://www.openarchives.org/sfc/sfc_oams.htm">
  </List-Meta-Formats>
```

List Partitions

Verb: **`List-Partitions`**
Version: 2.0
Fixed args: none
Keyword args: none
Return MIME type: `text/xml`
Return Status Codes: 200, 400

Return a structured list of the administrator-defined partitions for this archive. The list contains the hierarchy of partitions and sub-partitions. For each partition, both the short name and long description is returned. Depending on the policy for a particular archive, a record may be a member of more than one partition.

Example Request:

```
/Dienst/Repository/2.0/List-Partitions
```

Example Response:

The following response indicates a partition hierarchy with two top level partitions - `Oceanside` and `ValleyView` - each with partitions hierarchies within them.

```
<?xml version="1.0" encoding="UTF-8"?>
  <List-Partitions version="2.0">
    <partition name="Oceanside">
      <display>Oceanside University of Nebraska</display>
      <partition name="CompEnt">
        <display>Department of Computational Entomology</display>
      </partition>
      <partition name="MetPhen">
        <display>Department of Metaphysical Phenomenology</display>
      </partition>
    </partition>
    <partition name="ValleyView">
      <display>Valley View University of Florida</display>
      <partition name="Fren">
        <display>Department of Frenetics</display>
      </partition>
      <partition name="Hist">
        <display>Department of Histrionics</display>
      </partition>
    </partition>
  </List-Partitions>
```

List Metadata Formats available for a Record

Verb: **`Structure`**
Version: 2.0
Fixed args: `fullID`
Keyword args: `view`
Return MIME type: `text/xml`
Return Status Codes: 200, 400, 404

This verb returns a structured response that describes the metadata formats available for a record. A client may use this information as the basis for metadata requests using the `Disseminate` verb.

There is one required keyword argument that can only take one value (the same verb in the full Dienst protocol has more keyword arguments that take more values):

- `view` with the single value `#`.

Example Request:

```
/Dienst/Repository/2.0/Structure/handlecorp/970101?view=%23
```

Example Response:

```
<?xml version="1.0" encoding="UTF-8"?>
  <Structure version="2.0">
    <meta-formats>
      <rfc1807 />
      <dc />
    </meta-formats>
  </Structure>
```

This response says that the record can disseminate two metadata formats `rfc1807` and `dc` (Dublin Core).

Appendix - DTDs for Messages

The Open Archives Metadata Set

Introduction

The Santa Fe Convention provides recommendations for interoperability among archives. Archives provide access to *records*. The word *document* is purposely avoided and the notion of a *record* is purposely imprecise. Some archives may just provide access to metadata, others may also provide access to metadata and full content in some form, others may provide other services associated with the metadata and content such as access to the full content in various manifestations (formats) or structural decompositions (e.g., individual pages, chapters, and the like).

This document describes the elements of the Open Archives Metadata Set (oams). The semantics of this set has purposely been kept simple in the interest of easy creation and widest applicability. 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.

Notes on the remainder of this document:

- The transfer syntax of the oams is XML. A DTD is given at the end of this document.
- The semantics of the oams could be expressed using Dublin Core Element Set, with some qualification of those elements. Where the semantics of an element in the oams exactly matches that of one in the Dublin Core Metadata Element Set, the definition of the Dublin Core element has been used.
- Except where noted, values for elements are unformatted strings.
- All dates in oams are encoded using the "Complete date" variant of ISO8601. This format is `CCYY-MM-DD` where `CC` is the century, `YY` is the year, `MM` is the month of the year between 01 (January) and 12 (December), and `DD` is the day of the month between 01 and 28 or 29 or 30 or 31, depending on length of month and whether it is a leap year.
- Elements that are mandatory are annotated with a `[M]`.
- Elements that are optional are annotated with a `[O]`.
- Elements that are repeatable are annotated with a `[R]`. Note that the fact that an element may be repeated implies that multiple values should not be associated with a single element (e.g., associating multiple authors with a single Author tag).

Description of the semantics of the Open Archives Metadata Set

Title [M]

A name given to the record.

Date of Accession [M]

The date when the record was entered into the archive. It is assumed that in most cases this date will be created automatically by the archive rather than entered by a human user.

Display ID [O] [R]

A URL (Universal Resource Location) identifying a human readable page that provides access to the possible manifestations (e.g., PostScript, TeX) of the record. For archives that have only one manifestation per record, this URL may point to that single manifestation.

Full ID [M]

The full identifier for a record in an archive. This full identifier is the concatenation of the following components:

I. A unique archive identifier consisting only of alphanumerical characters [`a-z`, `A-Z`, `0-9`]. Registration of this identifier is part of the Open Archives registration process for data providers, described in Step 6 of the core document of the Santa Fe Convention.

II. Any printable non-alphanumeric character that will act as a delimiter (e.g., `/ : #`)

III. An identifier for the record that is unique within the archive.

The combination of these components produces a globally unique full identifier for each record in the nature of a URN. An example of a Full ID is `archive11/xxx4`.

Author [M] [R]

The author or corporate author who is responsible for creating the intellectual content of the record. Each author may also have an optional institution affiliation.

Abstract [O]

Text summarizing the contents of the record.

Subject [O] [R]

The topic of the content of the record expressed as keywords, key phrases or classification codes.

Comment [O] [R]

A free-text value that contains information outside the scope of other defined elements that adds to the discoverability of the record.

Date for Discovery [O] [R]

A date relevant to the record that may aid the user trying to find the document. A common example of such a date would be an original publication date of a record that was placed in an archive at a later time (i.e., its date of accession is later than its date of publication).

XML DTD for the Open Archives Metadata Set

The plain text DTD file can be retrieved here. The oams DTD can be embedded in a larger DTD.

```
<!-- Open Archives Metadata Set (oams) -->
<!-- This DTD can be used to represent the elements of the
Open Archives Metadata Set-->
<!-- Version 0.2, Mark Doyle Dec 27, 1999 -->
<!-- Dates are to be in encoded using the "Complete Date" variant of
ISO8601-->
<!ENTITY % doctype "oams">
<!ELEMENT %doctype; (title, accession, displayId*, fullId, author+,
abstract?,subject*,comment*,discovery)>
<!ELEMENT title (#PCDATA)>
<!ELEMENT accession EMPTY>
<!ATTLIST accession date CDATA #REQUIRED>
<!ELEMENT displayId (#PCDATA)>
<!ELEMENT fullId (#PCDATA)>
<!ELEMENT author (name,organization*)>
<!ELEMENT name (#PCDATA)>
<!ELEMENT organization (#PCDATA)>
<!ELEMENT abstract (#PCDATA)>
<!ELEMENT subject (#PCDATA)>
<!ELEMENT comment (#PCDATA)>
<!ELEMENT discovery EMPTY>
<!ATTLIST discovery date CDATA #REQUIRED>
<!-- ENTITY sets - lifted from MathML DTD -->
<!-- ISO 9573-13 -->
<!ENTITY % ent-isoamsa SYSTEM "isoamsa.ent" >
%ent-isoamsa;
<!ENTITY % ent-isoamsb SYSTEM "isoamsb.ent" >
%ent-isoamsb;
<!ENTITY % ent-isoamsc SYSTEM "isoamsc.ent" >
%ent-isoamsc;
<!ENTITY % ent-isoamsn SYSTEM "isoamsn.ent" >
%ent-isoamsn;
<!ENTITY % ent-isoamso SYSTEM "isoamso.ent" >
%ent-isoamso;
<!ENTITY % ent-isoamsr SYSTEM "isoamsr.ent" >
%ent-isoamsr;
<!ENTITY % ent-isogrk3 SYSTEM "isogrk3.ent" >
%ent-isogrk3;
<!ENTITY % ent-isogrk4 SYSTEM "isogrk4.ent" >
%ent-isogrk4;
<!ENTITY % ent-isomfrk SYSTEM "isomfrk.ent" >
%ent-isomfrk;
<!ENTITY % ent-isomopf SYSTEM "isomopf.ent" >
%ent-isomopf;
<!ENTITY % ent-isomscr SYSTEM "isomscr.ent" >
%ent-isomscr;
<!ENTITY % ent-isotech SYSTEM "isotech.ent" >
%ent-isotech;
<!-- ISO 8879 -->
<!ENTITY % ent-isobox SYSTEM "isobox.ent" >
%ent-isobox;
<!ENTITY % ent-isocyr1 SYSTEM "isocyr1.ent" >
%ent-isocyr1;
<!ENTITY % ent-isocyr2 SYSTEM "isocyr2.ent" >
%ent-isocyr2;
<!ENTITY % ent-isodia SYSTEM "isodia.ent" >
%ent-isodia;
<!ENTITY % ent-isogrk1 SYSTEM "isogrk1.ent" >
%ent-isogrk1;
<!ENTITY % ent-isogrk2 SYSTEM "isogrk2.ent" >
%ent-isogrk2;
<!ENTITY % ent-isolat1 SYSTEM "isolat1.ent" >
%ent-isolat1;
<!ENTITY % ent-isolat2 SYSTEM "isolat2.ent" >
%ent-isolat2;
```

```
<!ENTITY % ent-isonum SYSTEM "isonum.ent" >
%ent-isonum;
<!ENTITY % ent-isopub SYSTEM "isopub.ent" >
%ent-isopub;
<!-- MathML aliases for characters defined above -->
<!ENTITY % ent-mmlalias SYSTEM "mmlalias.ent" >
%ent-mmlalias;
<!-- MathML new characters -->
<!ENTITY % ent-mmlextra SYSTEM "mmlextra.ent" >
%ent-mmlextra;
<!-- end of ENTITY sets -->
```

A sample record expressed according to the Open Archives Metadata Set XML DTD

The plain text sample record can be retrieved here.

```
<?xml version="1.0"?>
<!DOCTYPE oams SYSTEM "oams.dtd">
<oams xmlns="http://www.openarchives.org/sfc/sfc_oams.htm">
<title>Dilaton Contact Terms in the Bosonic and Heterotic
Strings</title>
<accession date="1992-01-30"/>
<displayId>http://arXiv.org/abs/hep-th/9201076</displayId>
<fullId>arXiv:hep-th/9201076</fullId>
<author><name>Mark Doyle</name><organization>Princeton
University</organization></author>
<abstract>Dilaton contact terms in the bosonic and heterotic strings are
examined following the recent work of Distler and Nelson on the bosonic and
semirigid strings. In the bosonic case dilaton two-point functions on the
sphere are calculated as a stepping stone to constructing a good coordinate
family for dilaton calculations on higher genus surfaces. It is found that
dilaton-dilaton contact terms are improperly normalized, suggesting that the
interpretation of the
dilaton as the first variation of string coupling breaks down when other
dilatons are present. It seems likely that this can be attributed to the
tachyon divergence found in Ref 1. For the heterotic case, it is found that
there is no tachyon divergence and that the dilaton contact terms are
properly normalized. Thus, a dilaton equation analogous to the one in
topological gravity is derived and the interpretation of the dilaton as the
string coupling constant goes through.</abstract>
<subject>High Energy Physics - Theory</subject>
<comment>Journal-ref: Nucl. Phys. B381 (1992) 158-200</comment>
<discovery date="1999-12-06"/>
</oams>
```

last updated February 11th 2000