First meeting of the Open Archives initiative

Initiative: Paul Ginsparg, Rick Luce, Herbert Van de Sompel

Meeting:

- **Location**: Santa Fe, New Mexico, US, October 21-22 1999
- **Sponsors**: Council on Library and Information Resources, the Digital Library Federation, the Scholarly Publishing and Academic Resources Coalition, Association of Research Libraries, the Research Library of the Los Alamos National Laboratory.
- **Meeting moderators**: Clifford Lynch & Don Waters.
- **Represented institutions/organizations**: American Physical Society, Andrew W. Mellon Foundation, Association of Research Libraries, California Institute of Technology, Coalition for Networked Information, Cornell University, Council on Library and Information Resources, Digital Library Federation, Harvard University, HighWire Press, Library of Congress, Los Alamos National Laboratory, Massachusetts Institute of Technology, NASA Langley, Old Dominion University, the Scholarly Publishing and Academic Resources Coalition, Stanford Linear Accelerator Center, University of California, University of Ghent, University of Southampton, University of Surrey, Vanderbilt University, Virginia Tech and Washington University.
- **Represented eprint-initiatives**: arXiv.org (=xxx), CogPrints, NDLTD, RePEc, EconWPA, NCSTRL, NTRS
- **Participants:** see seperate list

Executive Summary

The Open Archives initiative has been set up to create a forum to discuss and solve matters of interoperability between author self-archiving solutions, as a way to promote their global acceptance (see http://vole.lanl.gov/ups/ups.htm).

The first, largest and most important such archive is the Los Alamos National Laboratory (LANL) Physics Archive. Founded by Paul Ginsparg in 1991, LANL now houses over 100,000 papers, mirrored worldwide in 15 countries with over 50,000 users daily and still growing (see http://arXiv.org/cgi-bin/show_stats). Other disciplines and institutions have begun to create public research archives along the lines of LANL but what is needed are conventions that archives could adopt to ensure that they work together so that any paper in any of these archives could be found from anyone's desktop worldwide, as if it were all in one virtual public library.

The participants in the meeting were digital librarians and computer scientists specializing in archiving, metadata, and interoperability, and they included the founders of the principal public research archives that exist so far. The participants were diverse in their underlying motivations, but entirely unified in their objective of paving the way for universal public archiving of the scientific and scholarly research literature on the Web.

The group agreed on minimal technical requirements for archives. These will be published separately as the "Santa Fe Conventions" and, in the next six months, will be implemented in the existing archives.

Technical Summary

The first meeting concentrated on the creation of cross-archive end-user services. The aim was to try and identify general architectural and technical characteristics of archive solutions, that would facilitate the creation of such services. These characteristics could then be used as recommendations for existing and upcoming initiatives.

The meeting started off with a presentation and demonstration by a team consisting of Herbert Van de Sompel (University of Ghent and Los Alamos National Laboratory), Michael Nelson (NASA Langley and Old Dominion University) and Thomas Krichel (University of Surrey and RePEc initiative). This group had built an experimental end-user service providing access to

data originating from main archive initiatives (arXiv, RePEc, NCSTRL, NDLTD, NTRS). A variety of technologies were used in the project, including NCSTRL+ as the digital library service, intelligent objects called buckets as a means to store the archive metadata and the SFX linking solution as a means to interlink the eprint data with the traditional scholarly communication mechanism. The presentation identified problems that arose during the project, and discussion of those served to launch the meeting. This presentation was followed by position papers on interoperability issues presented by Carl Lagoze (Cornell University), Kurt Maly (Old Dominion University), Ed Fox (Virginia Tech) and Caroline Arms (Library of Congress).

Following the initial presentations, there was a panel discussion in which Paul Ginsparg (Los Alamos National Laboratory), Paul Gherman (Vanderbilt University), Eric Van de Velde (CalTech) and John Ober (University of California) expressed their opinion on the possible pros and cons of institutional versus discipline-oriented archive initiatives. The group concluded that many different archive initiatives were likely to emerge, with different conceptual, organizational and technical foundations. In order for such initiatives to successfully become part of the scholarly communication system, interoperability was seen as a crucial factor.

The group agreed that interoperability hinges on a fundamental distinction between the archive-functions, which include data-collection and maintenance and end-user functions, like the cross-system search and linking prototype service described in the opening session. Although archive initiatives can implement their own end-user services, it is essential that the archives remain "open" in order to allow others to equally create such services. This concept was formalized in the distinction between providers of data (the archive initiatives) and implementers of data services (the initiatives that want to create end-user services for archive initiatives). Stimulated by a presentation by Thomas Krichel, the group agreed that an essential feature of the Santa Fe Conventions would be that providers of data use a standard mechanism to state the conditions under which their datasets can be used by implementers of data services. Similarly, the implementers of data services could describe the use they make of archive data.

This organizational argument was followed by a discussion on the technicalities of creating end-user services for data originating from different archives. The group recognized that there are basically two ways to implement these: a distributed searching approach and a harvesting approach. The former would require archives to implement a joint distributed search protocol, which is not considered to be a low-entry requirement. Moreover, the technical experts recognized that there are important problems of scale when implementing such distributed search solutions, in light of the possible emergence of thousands of institutional and/or subject-oriented archives worldwide. As such, the group decided this was not a realistic approach at this point in time. Therefore, as in the experimental project presented at the beginning of the meeting, a harvesting solution was proposed. Such a harvesting solution would allow trusted parties - the ones that subscribe to the Santa Fe Conventions - to selectively collect data from different archives. It was identified that such a technique requires an understanding regarding:

- Protocols to selectively harvest data;
- Criteria that can be used to selectively harvest data;
- Metadata formats that are used by archive solutions to respond to harvesting requests.

It was recognized that providers of data could describe the details of these interfaces in standard ways thus enabling implementers of data to create archive-specific harvesters. Still, the group decided to go one step further and to highly recommend the following:

- Protocols to selectively harvest data: implementation of part of the Dienst protocol in order to achieve a uniform way to poll an archive for its logical division(s) (subarchives) and to selectively harvest data from these divisions.
- Criteria that can be used to selectively harvest data: there should at least be support for a bulk harvest of all data from an archive, as well as a mechanism to harvest based on accession date. Other harvesting criteria that were thought to be important included author affiliation, subject, publication type.
- Metadata formats that are used by archive solutions to respond to harvesting requests: It is recognized that archives will use (an) internal metadata format(s) best suited to deal with the material to be described. Still, the group decided to propose a minimal Dublin Core compliant metadata set, called the Santa Fe Set, that should be made available by all archives. It is desirable that archives are able to respond to harvesting requests with data delivered in both the internal metadata format as in the Santa Fe Set format.

The representatives of existing archive initiatives at the meeting as well as those from institutions that are in the process of setting up archive initiatives agreed to comply to those guidelines. The Dienst protocol will be enhanced to allow for the functions mentioned above and a minimal Dienst release facilitating the process of making an archive compliant to the required aspects of Dienst will be made available. A transport format for MARC-formatted metadata will be proposed, as well as an XML DTD for

the description of the Santa Fe Set. The recommendations will be extensively documented on a Web site. Adoption of the recommendations will be promoted worldwide.

The way forward

- The minimal Dienst protocol set will be implemented for all archives that were represented at the meeting. This will allow for a first round of experimentation with the creation of end-user services layered over existing archives.
- The group identified the urgent need to discuss the mechanisms used to submit material to archives.
- Paul Ginsparg suggested that a next meeting should be held in Europe, in the first quarter of next year.
- It was also thought to be important to have a presentation and/or workshop on the Open Archives Initiative at the ACM 2000 Conference on Digital Libraries as well as at the European ECDLC.
- The experimental, non-productional prototype that was presented at the meeting will temporarily be available for exploration at the beginning of November 1999 at http://ups.cs.odu.edu . The representatives of Old Dominion University, the Research Library of the Los Alamos National Laboratory and the University of Ghent expressed their interest in continuing this prototyping work.
- The UPS Initiative has been renamed. From now on it is the Open Archives initiative.

October 29th 1999

get in touch with the Open Archives initiative by contacting
herbert.vandesompel@rug.ac.be