The open archives initiative

Forging a path toward interoperable author self-archiving systems

by Richard E. Luce

Efforts aimed at giving authors control over the communication and distribution of their work, in the form of electronic author self-archiving systems, are gaining ground. The Universal Preprint Service (UPS) initiative, now known as the Open Archives Initiative, is one such effort that has been widely publicized and often misreported in recent accounts. The initiative's goals are to develop a framework for a "universal e-print archive" and to establish interoperability standards supporting the search and retrieval of e-print papers from all disciplines. The hope is to catalyze progress towards new scholarly publishing models over the next five-to-ten years.

Background

Scholarly communication has long suffered from the lag time between summarizing research results in a publishable article and the formal publication of the article. In certain areas of scholarly activity, electronic preprint archives have become an established medium to communicate non peer-reviewed results of ongoing research quickly.

The trend began in high energy physics in 1991 and, since then, the centralized xxx preprint archive founded by Paul Ginsparg at Los Alamos National Laboratory has become a global repository for research in physics. xxx (also known as arXiv.org) houses more than 122,000 papers, and is mirrored worldwide in 15 countries, with over 60,000 users daily. xxx has now expanded to incorporate mathematics, non-linear sciences and computer science.

Similar efforts are taking place in other disciplines. CogPrints is modeled on xxx and focuses mainly on a collection of papers in cognitive science, psychology, neurology, linguistics, and related fields. NCSTRL (Networked Computer Science Technical Reports) is a similar initiative, providing a point of access for technical reports in computer science that are either submitted to the CoRR (Computing Research Repository), a part of xxx, or to decentralized departmental archives that cooperate in the initiative. Archives in the NCSTRL initiative share the Dienst protocol, which enables the creation of library-like services that support searching and browsing the archive.

Along the same lines, the RePEc (Research Papers in Economics) initiative provides authors with the option to submit working papers to a departmental archive or, if one does not exist, to the EconWPA archive at Washington University.

The NDLTD project (Networked Digital Library of Theses and Dissertations) aims at building a digital library of electronic theses and dissertations (ETD) authored by students of member institutions. NDLTD addresses issues such as the creation of a workflow to

About the author

Richard E. Luce is library director and Library Without Walls project leader of the Research Library at Los Alamos National Laboratory, e-mail: rick.luce@lanl.gov

submit ETDs, the development of an XML Document Type Definition for ETDs and the support of a digital library for ETDs. Recently, NIH has expressed a strong interest in the establishment of PubMed Central, an e-print initiative for biology described in the first column in this series *(C&RL News,* January 2000). All of these preprint initiatives endeavor to optimize scholarly communication by overcoming the barriers—financial, legal, etc.—which the traditional framework has established.

While other disciplines and institutions have begun to create public research archives along the lines pioneered at Los Alamos, what is needed are conventions that archives can adopt to ensure that they work together. Ideally, any scholar should be able to find any paper in any of these preprint or e-print archives from any desktop worldwide, as if they were all in one virtual public library. The information industry is slowly beginning to understand the potential of the preprint concept, regarding it either as an opportunity for collaboration, as a challenge, or as a threat.

Taking the first step—UPS Initiative

In April 1999, a call for participation was put out to existing e-print systems to mobilize a core technical group to work towards achieving a universal service for non peer-reviewed scholarly literature. Such a universal service is considered as the fundamental and free layer of scholarly information, on top of which both free and commercial services could flourish.

Paul Ginsparg, Herbert Van de Sompel and I, from Los Alamos National Laboratory, initiated the UPS Initiative call for participation. We believed that important steps towards the establishment of such a universal service could be taken by identifying or creating interoperable technologies and frameworks for the dissemination of author self-archived documents (termed e-prints). The driving forces behind the initiative were the perception that many years of theoretical discourse have resulted in few fundamental methodological changes, and our hope that more-rapid progress could be catalyzed by a consortium of interested parties focusing directly on the relevant technological issues.

The UPS meeting was held in Santa Fe, New Mexico on October 21–22. 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. 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. Sponsorship for the meeting was obtained from the Council on Library and Information Resources; the Digital Library Federation; SPARC; ARL; and the Research Library at the Los Alamos National Laboratory.

A set of objectives was outlined for the meeting. These objectives supported the development of solutions to some of the purely technical obstacles to a more-effective electronic scholarly communication system and centered around the following concepts:

1. stimulating the adoption of the preprint concept in all areas of scholarly research;
2. integrating preprint services into the scholarly document system of scholarly journals, A&I services and libraries;
3. creating search and retrieval functionality for preprint archives that can be simultaneously useful for discipline-specific, cross-disciplinary, inter-institutional and intra-institutional purposes;
4. developing user-friendly systems, i.e., along the lines of established search and retrieval methods; and
5. including the full range of meta-data, full-text, and citation data.

The group agreed on a set of minimal technical requirements for archives. These will be published separately as the "Santa Fe Conventions" and, during the next six months,

About the editors

Ivy Anderson is coordinator for Digital Acquisitions at Harvard University, e-mail: ivy_anderson@harvard.edu; Gail McMillan is head of the Digital Library and Archives (formerly the Scholarly Communications Project) at Virginia Tech University, e-mail: gailmac@vt.edu; Ann C. Schaffner is associate university librarian for Research Services, Instruction & Planning at Brandeis University, e-mail: schaffne@brandeis.edu

will be implemented in the existing archives.

Technical summary

All the participants agreed that scientific papers should be freely accessible to the public, although individual participants differed on specifics, such as how to handle non-peer-reviewed material. The first meeting concentrated on the creation of cross-archive end-user services. The aim was to identify general architectural and technical characteristics of archive solutions that would facilitate the creation of such services. These characteristics could then be recommended for existing and upcoming initiatives.

The meeting began with a presentation and demonstration by a team consisting of Herbert Van de Sompel, 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 existing archive initiatives. The presentation identified problems that arose during the project, and discussion of these problems served to launch the meeting discussions.

Participants concluded that many different archive initiatives were likely to emerge, with different conceptual, organizational, and technical foundations. For such initiatives to become part of the scholarly communication system, interoperability was essential.

Consensus was reached 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" to allow others to create such services.

A discussion on the technicalities of creating end-user services for data originating from different archives followed. 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 would be difficult. Moreover, 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. The group agreed that this was not a realistic approach at this time, and that a harvesting solution was more appropriate. Such a harvesting solution would allow trusted parties—the ones that subscribe to the Santa Fe Conventions—to collect data selectively from different archives. The conventions propose adoption of portions of the Dienst protocol for the harvesting of data and a minimal Dublin Core compliant metadata set, called the *Santa Fe Set*, which should be made available by all archives to respond to harvesting requests.

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 with 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 Document Type Definition for the description of the Santa Fe Set. The recommendations will be extensively documented on a Web site and adoption of the recommendations will be promoted worldwide.

The path forward

The Open Archives initiative has created a forum to discuss and solve technical matters of interoperability between author self-archiving solutions, as a way to promote their global acceptance (see http://www.openarchives.org).

(continued on page 202)

URLs of Eprint archives

- xxx preprint archive at Los Alamos: http://xxx.lanl.org
- CogPrints: http://cogprints.soton.ac.uk
- NCSTRL: http://www.ncstrl.org
- EconWPA: http://econwpa.wustl.edu
- NDLTD: http://www.ndltd.org


No Limits.
The quantity of information available is virtually limitless. Shouldn't you expect the same in quality and accessibility from your information services provider? Our comprehensive subscription management services provide access to more than 260,000 title listings in a variety of formats. A sophisticated electronic journal service and a multitude of full text online reference databases offer accurate, up-to-date content from thousands of the world's most respected sources. And we link it all electronically to provide you with seamless access to full text, no matter where your research begins.
At EBSCO, we're not just pushing the limits — we're surpassing them.
EBSCO
INFORMATION SERVICES
www.ebsco.com
P.O. Box 1943 • Birmingham, AL 35201-1943 • (205) 991-6600

biographical entries and a description of the discovery that earned the Nobel Prize. *Access:* http://www.slac.stanford.edu/library/nobel.html.

• **Contributions of 20th Century Women to Physics.** An archive of more than 80 citations of women who contributed to physics in the 20th century created by the UCLA Department of Physics. *Access:* http://www.physics.ucla.edu/~cwp/.

History of Physics

• **Center for the History of Physics.** This center, a section of the American Institute of Physics, has a mission to preserve and make known the history of modern physics and allied fields, including astronomy, geophysics, optics, and the like. *Access:* http://www.aip.org/history/.

• **History of Physics Group.** The Institute of Physics established this group in 1984 to secure the written, oral, and instrumental record of British physics for posterity and to explore ways in which history can be used more effectively in the understanding, teaching, and general communication of physics. *Access:* http://www.iop.org/IOP/Groups/HP/.

Discussion groups

A key mailing list for physics librarians is PAMnet, the discussion list for the Physics, Astronomy and Mathematics (PAM) Division of the Special Libraries Association. The purpose of PAMnet is to provide a forum for the discussion of library and information resource issues relevant to physics, astronomy, and mathematics. PAMnet may be used to seek help with reference questions and in obtaining materials, but only when those materials are not available through a library's normal ILL or document delivery suppliers or when timing is critical. The list is open to non-PAM division members. To subscribe, contact David Stern, the list owner, at david.e.stern@yale.edu.

Other mailing lists related to physics can be found in the Physics section of "The Directory of Scholarly and Professional E-Conferences," maintained by Diane Kovacs. *Access:* http://www.n2h2.com/KOVACS/S0106.html.

There is a physics hierarchy of discussion groups in Usenet. These include sci.physics, sci.physics.relativity, and sci.physics.research. ■

"Scholarly Comm." cont. from page 186)

Agreement was reached on the following to pave a path 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;

• there is an urgent need to discuss the mechanisms used to submit material to archives;

• it is important to have presentations and/or workshops at upcoming digital library conferences;

• the experimental, non-production prototype that was presented at the meeting will temporarily be available for exploration at http://ups.cs.odu.edu. The representatives of Old Dominion University, Los Alamos National Laboratory Research Library, and the University of Ghent expressed their interest in continuing this prototyping work; and

• the UPS Initiative has been renamed. It is now referred to as the Open Archives Initiative.

Some issues and questions

The initiative discussed above raises several social issues concerning scholarly communication. Among the issues of relevance to academic and research institutions are the following:

• Will the institution provide or support a departmental or institutional e-print archive of authors affiliated with the institution? If so, will it adopt or incorporate the Santa Fe protocols to gain wider exposure and interoperability?

• How will research libraries package and deliver access to e-print literature?

• With the resolution of e-print archive interoperability technical issues, what will be the process of resolving the social issues connected with tenure and publishing?

While it is not the intent of the Open Archives Initiative to deal with those social issues, their resolution will be an important factor in determining how quickly the paradigm for scholarly communication will change. At our meeting in October, we tried to lay the groundwork for technical standards that will support new models of scholarly publishing. ■