The Open Archives Initiative: Interoperable, Interdisciplinary Author Self-archiving Comes of Age

Richard E. Luce
Research Library Director & Library Without Walls Project Leader,
Los Alamos National Laboratory Research Library

Serials Review and NASIG Conference Proceedings, San Diego, June 22-25 2000.

Abstract

Author self-archiving systems, emerging from successful experiments with preprint servers, have emerged in a variety of fields. The Open Archives initiative was organized to create a forum to solve interoperability issues between author self-archiving solutions, as a way to promote their global acceptance. The initiative seeks to develop a framework for a "universal e-print archive" that establishes interoperability standards supporting the search and retrieval of e-print papers from all disciplines. The Santa Fe conventions were developed to ensure these archives work together so that any paper in any of these archives can be found from anyone's desktop worldwide, as if it were all in one virtual public library.

A revolution in the scholarly communication system is brewing with the goal of returning choice back to authors. Efforts to give authors control over the communication and distribution of their work, in the form of electronic author self-archiving systems, are gaining ground. Author self-archiving allows authors to deposit their papers or preliminary drafts into an archive and thereby speed up the communication process. Submittal for publication and peer review follows later, if desired by the author. These archiving alternatives, typically organized by subject domain or organizational entity, are growing and rapidly gaining acceptance.

Until very recently, however, these developments have been relatively uncoordinated and somewhat isolated "islands" of information. The obvious challenge for libraries and researchers is the question of locating relevant content among heterogeneous and highly variable systems, or simply put the ability to interoperate on these systems as one virtual collection.

The Open Archives initiative represents an attempt to develop a framework for a "universal e-print archive" that establishes interoperability standards supporting the search and retrieval of e-print papers from all disciplines (1). At the most basic level interoperability is defined as "the capacity of a user to treat multiple digital library collections as one" (2) and it is widely considered a key digital library challenge.

While author self-archiving systems today are much broader in format than preprint archives, nonetheless the genesis of this movement comes from the preprint experience. What lessons can we learn from the successful preprint servers, and what are the implications for scholarly communication?

Los Alamos E-print Archive

The first and most important preprint server and archive is Los Alamos National Laboratory's physics preprint archive www.arXiv.org, which expanded to support mathematics, nonlinear sciences and computer science (formerly known as xxx.lanl.gov). Created by Paul Ginsparg in 1991 to speed the delivery of high energy physics preprints, the arXiv has become the global repository for research in physics. The arXiv contains over 134,000 papers and receives about 2,500 new author submissions monthly. Mirrored in 15 countries, it receives constant and heavy usage, supporting an average of 120,000 daily connections.

Many players in the value chain advanced arguments to support the contention that the preprint model would not expand outside physics to other disciplines. We now know that is not the case. After sorting out obviously self-serving rationale to protect the status-quo, some concerns are quite valid (3). However, this model was not intended to be the all-encompassing solution. Rather than focusing on dissimilarities in cultural communication in different fields of research, it is more instructive to note cross-disciplinary similarities. Speed, cost and value chain issues are not limited to the physics community and those issues are among the factors that are driving experimentation with author self-archiving systems.

Other E-print Efforts

Similar efforts in other disciplines are noteworthy since the Open Archives initiative seeks to address interoperability among these systems and others in early stages of development. Many began as informal mechanisms to disseminate either preliminary results or grey literature. A number of these have evolved into essential vehicles to communicate results to colleagues in a given domain.

- CogPrints is modeled on the arXiv and focuses primarily on a collection of papers in cognitive science, psychology, neurology, linguistics, and related fields.
- Archives in the NCSTRL (Networked Computer Science Technical Reports) provide access to technical reports in computer sciences from over 100 institutions worldwide through a single interface (4). The initiative uses the Dienst protocol, which enables the creation of library-like services that support searching and browsing the archive.
- The RePEc initiative (Research Papers in Economics) 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.
- NDLTD aims at building a digital library of electronic theses and dissertations (ETD) authored by students of member institutions. It contributes a useful and unique area of "grey" literature that otherwise would be available only through a commercial service or directly from each university.
- NASA Technical Reports Server (NTRS) is a gateway to 20 different U.S. government-based technical report servers that contain three to four million abstracts and more than 100,000 full-text reports.
- Clinical Medicine Netprints, launched recently by the British Medical Journal and HighWire Press is an e-print site for studies, research, and articles in Clinical Medicine (5).

- Recently, NIH has expressed a strong interest in the establishment of an e-print initiative for biology. The NIH e-biomed proposal (6) for research reports in the life sciences has received a significant attention and debate. PubMed Central, though representative of a more traditional approach, provides barrier-free access to primary reports in the life sciences. (7)

E-print Lessons

What drives the rapid adoption of these systems, still in relative infancy in terms of their development? From a market perspective, the old paradigm for scholarly communication was not adequately meeting the primary needs of its customers, in this instance the scientific community. Clearly the traditional scholarly model of communication, with its reliance on formally published journals, is facing significant challenges. Although it is not within the scope of this paper to explore those challenges, three key factors are critical to understand the adoption of a new model:

- **Speed** - In a world shaped by the Internet, scientists now have access to a medium that supports rapid communication and sharing of research results. Today scientific research in most fields is moving faster than ever. Rapid communication drove initial efforts to launch xxx.lanl.gov in 1991 and it drives the adoption of alternatives in other fields today.

- **Financial Instability** - The imbalance between double digit pricing increases and relatively flat library budgets has created a well-known financial crisis for research libraries. It has also negatively impacted the author scientist who typically pays for this imbalance with institutional overhead taxes taken out of his or her research funding.

- **Inefficient Value Chain** - A primary motivation of the author/researcher is the accreditation and communication of results to one's peers. The current value chain for formal publication is very long, with several layers between author and reader. (Eg., author, editor, peer review, primary publisher, secondary publisher, distributor, library, and reader). It is reasonable to conclude that this chain, with various inefficiencies, is not sustainable in its current form.

To condense these points into an equation, we could state that significantly increasing the speed of communication, coupled with radically lower costs and close proximity between author and reader equals a formula for success. All the e-print initiatives share the same goal, the optimization of scholarly communication by overcoming the barriers imposed by the traditional framework.

A final general observation, and by no means the least important issue, is the lack of direct user involvement in the large fraction of currently available formal communication systems. Relatively very few practicing scientists are involved in the scholarly communication debate, much less in the design and implementation of new systems. That reality is ironic given scientists feed the scientific publication process as authors, as well as actively interact with the formal system on a daily basis as readers, referees, editors, conference organizers, etc.

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 are interoperable. Ideally, any paper in any of these preprint or e-print archives should be able to be found from anyone's desktop worldwide, as if it were all in one virtual public library.

Taking the First Steps

In April 1999 a call for participation for a Universal Preprint System (UPS) was put out to existing e-print systems. This was intended to mobilize a core technical group to work towards achieving a *universal service for non peer-reviewed scholarly literature* (8). Such a universal service is considered as the fundamental and free layer of scholarly information, on top of which both free and commercial services can be established. The goal is to catalyze progress in new scholarly publishing models over the next five to ten years.

The call for participation was based on the premise that important steps towards the establishment of 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 force behind the initiative is the perception that many years of theoretical discourse have resulted in few fundamental methodological changes, and the hope that more-rapid progress could be catalyzed by a consortium of interested parties' focusing directly on the relevant technological issues.

The first UPS meeting was held in Santa Fe, N.M. on October 21-22, 1999. The participants 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 unified around the objective of paving the way for universal public archiving of the scientific and scholarly research literature on the Web.

A second meeting connected with the Open Archives initiative was held on June 3 in San Antonio, Texas. The intent was to ratify, solidify, and expand on previous agreements (9). At the Second OAi meeting, 43 people assembled from 5 countries. As of the meeting date, there were 6 conforming archives with content available for harvesting. The third OAi meeting will be held in Lisbon, Portugal, on September 21, 2000, in conjunction with the September 18-20 activities of ECDL'2000 (10). Coordination will be provided by an emerging OAi steering committee to support the work of the initiative.

Technical Summary

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

Participants 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 essential. The initiative aims to support archives, both those focused on e-prints and those representing a wide variety of other content types. Version 2 of the OAi specifications and a number of conforming implementations, including in PERL and Java, will be available so that archives can participate easily in

OAi. More detailed descriptions of the meeting (11, 12, 13) and the prototype system demonstrated at the Santa Fe meeting (14) has been published.

Interoperability

The Santa Fe Convention of the Open Archives initiative represents a pragmatic and incremental approach towards interoperability. 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" in order to allow others to equally create such services.

Essentially, there are two ways to implement end-user services for data originating from different archives: either a distributed searching approach or a harvesting approach. The former would require archives to implement a joint distributed search protocol, which is difficult. Moreover, there are important problems of scale when implementing such distributed search solutions, given the possible emergence of thousands of institutional and/or subject-oriented archives worldwide. Thus the harvesting solution was considered more appropriate. The harvesting approach allows trusted parties -- the ones that subscribe to the Santa Fe Conventions -- to selectively collect data 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 mechanisms for establishing this interoperability are described in full detail in the Santa Fe Convention (15). The Santa Fe Convention presents a technical and organizational framework designed to facilitate the discovery of content stored in distributed e-print archives. It makes easy-to-implement technical recommendations for archives that – when implemented – will allow data from e-print archives to become widely available via its inclusion in a variety of end-user services. Authors can make electronic documents available to a global audience by submitting them to e-print archives. Interoperability is achieved by use of the following methods:

1. Specifying the protocol to harvest metadata from participating archives;
2. Specifying what criteria will be supported to selectively harvest desired metadata; and
3. Use of a common metadata format for supplier archives to use when responding 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.

Beyond the basic goal of accessing multiple archives as one collection, the term interoperability also implies other capabilities that use discovery tools on virtual collections (16). At a high level, value-added services that support discovery and personal alerting, rich dynamic linking, reviews and notation, and metrics that feed

recommendation systems and citation analysis can be envisioned. Rather than requiring each archive to create and support such capabilities, the Santa Fe Convention adopts the free market system. Any service provider is free and able to develop enhanced capabilities, allowing a competitive market to drive improvements.

Conclusion

The major achievement of the Santa Fe meeting is the agreement among a core group of pioneers and implementers to use cooperation to facilitate the further development of a broad e-print community. Serious consideration has been given to lowering the financial barriers that might preclude new participants in an effort to build momentum and wide adoption of publishing alternatives.

With the growing use of e-print archives, we are witnessing a transition phase from the old model of formal scholarly communication to a rapidly evolving hybrid. The new electronic medium provides an opportunity to reconsider many aspects of the current research communication process and the roles each of us play. It is an opportune time to experiment and rethink the assumptions that underlie our systems. Ginsparg believes "we should take advantage of this opportunity to map out the ideal research communication medium of the future. It is crucial that the researchers, who play a privileged role in this as both providers and consumers of the information, not only be heard but be given the strongest voice. In particular, we need to dislodge definitively the curiously prevalent notion that the future electronic medium will strictly duplicate, inadequacy for inadequacy, the current print medium"(17). I submit it is equally crucial for librarians to not only chime in with strong voices, but to rethink our vision and roles. And after having done that, to provide creative leadership during this transition phase.

Notes

1. Ginsparg, Paul, Rick Luce, and Herbert Van de Sompel. The Open Archives initiative. July 1999. [www.openarchives.org/]

2. Tennant, Roy. Library Journal. November 15, 1997, p. 31-32.

3. Boyce, Peter B. For Better or Worse: Preprint Servers Are Here to Stay. College & Research Libraries News: May 2000, p 404-407.

4. Leiner, B.M. 1998. The NCSTRL Approach to Open Architecture for the Confederated Digital Library. D-Lib Magazine 5, no. 12.

5. Delhamothe, Tony and others. 1999. Netprints: the Next Phase in the Evolution of Biomedical Publishing. British Medical Journal, 319: 1515-6.

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

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

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

9. See: http://purl.org/net/oaijune00/

10. See: http://purl.org/net/oaisept00

11. Luce, Richard E. "The Open Archives Initiative: Forging a Path Toward Interoperable Author Self-Archiving Systems". College & Research Libraries News: March 2000, p 184-186, 202.

12. Van de Sompel, Herbert and Carl Lagoze. "*The Santa Fe Convention of the Open Archives Initiative".* D-Lib Magazine. February 2000, Volume 6 Number 2. ISSN 1082-9873.

13. Fox, Edward A. "Open Archives initiative" D-Lib Magazine. June 2000 Volume 6, Number 6, ISSN 1082-9873.

14. Van de Sompel, Herbert, Thomas Krichel, Michael L. Nelson and Patrick Hochstenbach. The UPS Prototype: An Experimental End-User Service across E-Print Archives. D-Lib Magazine. February 2000, Volume 6 Number 2. ISSN 1082-9873.

15. For more information and additional references see the Web site at http://www.openarchives.org

16. Lagoze, Carl. 1999. Defining Collections in Distributed Digital Libraries, D-Lib Magazine 5, no. 11.

17. Ginsparg, Paul. First Steps Towards Electronic Research Communication, *Computers in Physics*, Vol.8, No.4, Jul/Aug 1994, p. 390.

Originally published in Serials Librarian, Volume 40, Numbers 1 / 2, 2001, pp. 173-182.

