Invited article

E-prints Intersect the Digital Library: Inside the Los Alamos arXiv

Richard E. Luce
Research Library Director & *Library Without Walls* Project Leader
Los Alamos National Laboratory
rick.luce@lanl.gov

Abstract

The e-print arXiv at the Los Alamos National Laboratory acts as a repository for electronic versions of papers in physics and mathematics, providing a rapid and convenient way for scientists to rapidly share their results with colleagues. Recently the arXiv was transferred to the Research Library, as part of its Library Without Walls. This article traces the development of the arXiv and examines some of the implications for libraries. Opportunities and challenges exist to integrate new forms of scholarly communication with newly developed digital library services offered by leading-edge libraries.

The E-Print Revolution

A far-reaching transformation of scholarly communication has thrived during the early stages of evolution, enabled by the computing and communications revolution. Developed in 1991 at the Los Alamos National Laboratory, the arXiv e-print server was the first widely successful automated electronic archive for research papers in physics and related disciplines: mathematics, nonlinear sciences, computational linguistics, and neuroscience. 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 (Holtkamp and Berg 2001). The online archive acts as a repository for electronic versions of papers, providing a convenient way for scientists to share their results with colleagues without having to wait for the article to appear in print. It is perhaps the best known example of the way the Internet has already changed the way scientists communicate.

Contextual Background

Los Alamos National Laboratory (LANL) is one of the largest multidisciplinary institutions in the world. Known for its world-class science, it is also the home of the e-print arXiv and a highly innovative Research Library. The Library's digital library project, the *Library Without Walls*, is recognized as a pioneering, state-of-the-art digital library (Pack and Pemberton 1999). While highly successful, the Library Without Walls initial efforts were concentrated on achieving a critical mass with digital collections and databases representing the formal scholarly scientific literature.

In June of 2000, administrative responsibility for daily operations of the Los Alamos arXiv was transferred from the Theoretical Physics Division to the Research Library. Added as a complement to the Library Without Walls, the move was intended to create and support synergies between new e-print advances and rapidly developing state-of-the-art digital library services. The move was significant on two levels. First, it represented a formalized commitment by the Laboratory regarding funding support of the arXiv. Secondly, it provided the Research Library with a unique opportunity to bridge the informal author self-archiving world with the formal scholarly communication systems in the digital library. Since the Research Library partners with research institutions around the world to enhance the advancement of information technology to support collaboration among researchers, developing a symbiotic relationship with the arXiv was a logical, and perhaps overdue, development.

Defining Preprints and E-prints

The term "preprint" often refers to a manuscript that has been peer-reviewed and is awaiting publication in a traditional journal. However, the term "preprint" also covers papers submitted for journal publication, but for which no publication decision has been reached, or even papers electronically posted for peer consideration and comment before submission for publication. In fact, preprints can also be documents that have not been submitted to any journal.

An "e-print" denotes self-archiving by the author. The American Physical Society notes that an e-print includes any electronic work circulated by the author outside of the traditional publishing environment (*APS News* 1998). E-prints may be unpublished works, preprints, or published works. Unlike the familiar paper preprints, the e-prints can be, and often are, updated by the author at any time, including after the peer-review process. In some subjects, where rapid transmission of knowledge is critical, electronic dissemination of preprints is an absolute necessity, with subsequent traditional publication becoming almost a formality (Langer 2000).

Addressing User Needs

A decade ago scholarly communication involved mail, fax, or more recently, FTP, and electronic mail. While traditional production and publication of documents requires a significant investment of time, materials, and money, placing a preprint or e-print on the World Wide Web involves no printing costs and practically no distribution costs. More importantly, preprint servers provide a convenient way for scientists to share their results with their colleagues very rapidly, in the form of individual articles. At this point, the vast majority of preprint servers contain scientific information. Fields in the humanities and social sciences have recently begun following the trend, but still lag significantly behind in terms of server repositories and papers.

Los Alamos arXiv

Paul Ginsparg, a Ph.D. physicist at LANL, grew frustrated with the slow existing paper preprint system and developed the first preprint archive in August 1991. Originally established to exchange papers in his field of high-energy theoretical physics, the arXiv (http://www.arxiv.org/) was formerly known as 'xxx' (xxx.lanl.gov). The Los Alamos arXiv is also known as the "Los Alamos e-print archive." At present the arXiv covers disciplines in physics, math, computer science, and non-linear systems. Author submissions are unrefereed and are a form of author self-archiving. There is no fee for either retrieval or submissions by users worldwide.

The archive is often contrasted with journals, but it was never intended to be a journal, and it was not intended to perform peer review. It is a repository for electronic versions of papers, providing a convenient way for scientists to share their results with colleagues. Since newly posted papers have not been peer reviewed, readers must use their own judgment regarding the trustworthiness of anything downloaded (*The Economist* 2000).

Evolution of the arXiv software

The software running the arXiv has evolved over the past nine years. The physics e-print archive was started with the *hep-th* (High Energy Physics - Theory) archive, supported with an e-mail interface. In 1992, the FTP interface was added. *Hep-ph* (High Energy Physics - Phenomenology) and *hep-lat* (High Energy Physics - Lattice) were added locally and *alg-geom* (Algebraic Geometry), *astro-ph* (Astrophysics) and *cond-mat* (Condensed Matter) were added remotely. By December 1993, a web interface was added. In November of 1994, data at some remote archives became mirrors. 1996 marked the growth of mirror networks and in June the web upload facility for author submissions was added.

Today the arXiv application code is comprised of roughly 30,000 lines of Perl, running under Linux with numerous other programs (TEX, ghostscript, tar, gnuzip). A small team of roughly four FTE supports the current activities of maintaining and rewriting the Perl code in modular fashion, supporting the mirror servers, and correcting author submissions. The application has been optimized to permit virtually any researcher with low-end network connections to access and download papers of interest. Very consciously, this provides a level playing field for researchers at different academic levels and different geographic locations.

Bottom Line: Is It Viable?

Two factors govern the ultimate viability of any scholarly communication system: (1) the input activity, or submission of content supplied by authors; and (2) usefulness, which is typically assessed via usage statistics. How does the arXiv fare in both of these dimensions?

Content Submissions

Authors make individual decisions whether to post their work on the arXiv. Since authors require readers, given the high usage of the arXiv for daily communication, sufficient incentive to post new papers exists. Over 155,000 papers have been submitted and posted, with over 30,000 new submissions posted in 2000. The system has experienced approximately linear growth in the submission rate, increasing by ~3,500 extra each year. Over 99% of the submissions are entirely automated and some journals now accept arXiv identifiers instead of requiring direct submission (e.g., American Physical Society: *Physical Review D*, Elsevier: *Physics Letters B*)

An analysis of the domain composition as a percentage of the total of all submissions (Aug '91 through Dec '00) follows:

High energy	Cond. matter	Astro-phys	math	Gen. relativity & quantum cosmology	Nuclear	Quant - phys	Non-Linear	Phys-Other	CS
37%	18%	17%	9%	5%	4%	3.5%	2.5%	2%	1%

Submission distribution according to e-mail domain of submitting author for 101,792 submissions received during the four year period 1 Jan 97 through 31 Dec 00 can be found at http://arXiv.org/Stats/au_all.html.

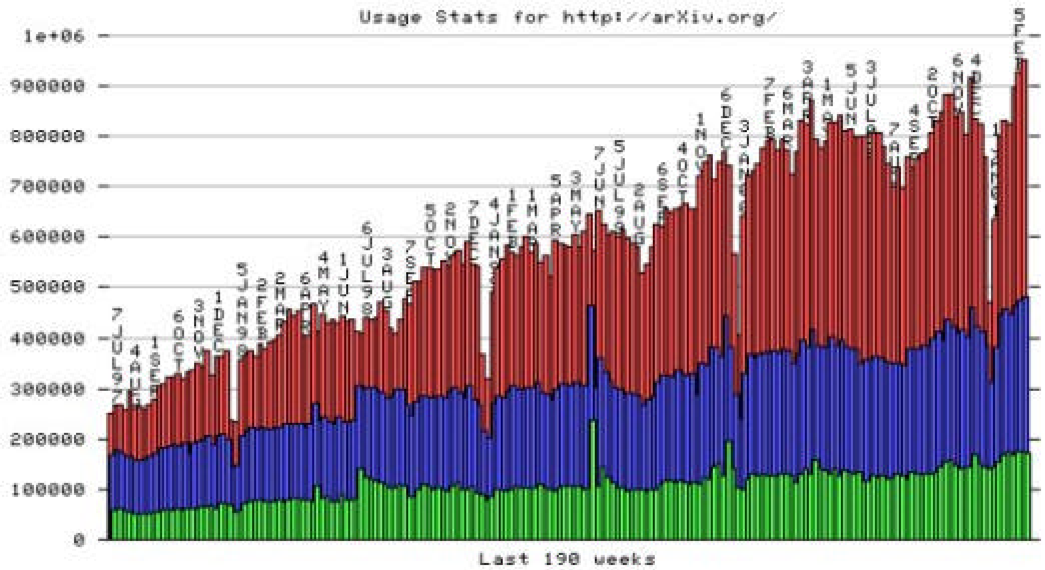
The table below shows the mode of submissions of papers via the web, electronic mail and FTP, from calendar '96 through calendar year '00:

Uploads	Late '96	'97	'98	'99	'00
Web	13%	21%	49%	60%	68%
E-mail	77%	67%	43%	34%	27%
FTP	10%	12%	8%	6%	5%

Authors can submit, replace and withdraw papers, add journal references, etc. (See: http://arxiv.org/help/). Manifestly bad papers in the archive are simply ignored. Those that are interesting and controversial attract instant attention. For better or worse, there is no time lag.

Usage Statistics

Use of the arXiv servers is very high by any measure. Attesting to the appeal in the international community, the system serves over 70,000 users in over 100 countries. An estimated 13 million papers were downloaded in 2000. The arXiv presently attracts from 110,00 to 130,000 visits daily.



Red - Number of **connections** in each week
Blue - Number of **hosts** connecting that week (divide by 10 for correct number)
Green - Number of **new** hosts that week (divide by 10)

The Slashdot Effect -- /.

The slashdot effect refers to the frenzy of download activity brought upon servers unexpectedly by announcements on slashdot.org which include a link to the original site. Slashdot is a very popular site for all kinds of "computing news" and /. or slashdot and to be "slashdotted" are idioms well established in the wider Linux and general computing community. Postings on slashdot have been known to crash or ground servers to a halt on multiple occasions. In fact, getting slashdotted turns out to be a good test of a server's and a network's capacity.

On Friday May 21, 1999 we received an article in the General Relativity and Quantum Cosmology section of arXiv titled "A 'warp drive' with more reasonable total energy requirements" from a

Belgian physicist. This is, despite its seemingly pure science fiction title (and well timed with the release of "the" new Star Trek episode at that time), a serious paper. Thus the paper was released and announced in the next mailing the following Sunday night/Monday morning. (See http://arXiv.org/abs/gr-qc/9905084.) Four days later on Thursday May 27, 02:14PM EDT slashdot posted a news story about this at http://slashdot.org/articles/99/05/27/1215204.shtml and almost instantly accesses to our server skyrocketed from noon to about 4pm that day, with a peak of over 16,000 hits in an hour. (See http://arXiv.org/Stats/9905/990527.html). Interestingly a great many readers only visited the abstract and didn't continue to download the full text, and a sizable fraction of downloads originated from physics departments all over the world. Since news on Slashdot is cycled away and moved to their archives within the same day, activity on the following day was back to almost normal levels. Our referrer logs show a few "late-comers," people catching up with yesterday's news, but there was no significant impact on the server. Perhaps just being exposed to the abstract is valuable for many casual readers. Nonetheless, the slashdot effect illustrates some of the problems inherent with suggestions that simply counting downloaded papers in this new environment can be used to gauge relative importance or impact.

Archiving -- Red Herring or Achilles Heel?

Archiving has been a much-discussed issue surrounding digitizing formal published works, with the effect of precluding progress in building digital collections by many research libraries.

One of the myths surrounding the arXiv is that the system is unstable as an archive because the e-print literature can come and go on an author's whim. Once a paper has been posted on the arXiv, it is given a date stamp and processed. Making changes results in a new version and the archive provides public access to previous versions of submitted papers. Therefore, even though the current version of a paper may be marked as withdrawn, previous versions can still be retrieved.

One of the problems that the e-print movement has had to confront is who should be responsible for the e-prints and how. Many supporters of e-prints want to see the traditional publishers removed from the role of caretakers of this scholarly communication genre, substituting either the professional societies or universities. However, this raises the major question of archiving. Historically, responsibility for archiving the scholarly literature has been the domain of libraries. Plagued with issues of cost, related to both content acquisition and the associated physical storage of print volumes, archiving print volumes has created a crisis for most research libraries.

Now with electronic archives, we replace those issues with the challenge of constantly migrating data formats and the associated enabling IT infrastructure for storage and access systems. Again, there are clearly cost issues surrounding the issue of constantly migrating technology and content formats. The archiving challenge becomes far more complex, however, when consideration is given to the question of preserving the robust environment that today's technology now supports. Increasingly, merely preserving the article itself cannot capture the value of an electronic article. Rather the value is in the associated contextual links, associated graphics, multi-media and connecting databases that have become intrinsic parts of modern scientific literature. Given this fact, in the very near term, the print versions of journals will not be the true archives.

One of the strategies of archiving is to physically distribute assets around to minimize the risk of a geographic disaster causing the irreplaceable loss of unique archive contents. In the e-print world, mirroring the data and access systems in several locations can mitigate this risk because a mirror image of both the data and the retrieval system is located in separate geographic sites. The Los Alamos arXiv is mirrored around the world in: Australia; Brazil; China; France; Germany, Israel; India; Italy; Japan; Russia; South Africa; South Korea; Spain; Taiwan; U.K.; U.S. (new APS mirror). Ironically, the Los Alamos arXiv is significantly better off considering this archiving dimension, as opposed to the current condition for a large fraction of formal publishers with electronic content located at one site and housed on one system.

More than Electronic Paper

Electronic e-prints do not simply represent what would appear in print journals. Indeed, for graphically dependent sciences, e-print publication on the web is preferable to paper journals because it offers numerous value-added elements, including multi-media, data sets, as well as linked references to other documents. Electronic preprints do not represent the only example of the technological impact of the greater efficiency and storage capacity of digital media. Now some new paper-based scholarly books have begun to omit printed bibliographies, instead referring readers to web sites. For example, Warren Siegel, a high energy physicist at SUNY-Stony Brook's C.N. Yang Institute for Theoretical Physics, is offering his own comprehensive textbook on quantum and classical field theory free of charge (*APS News* 2001). Entitled *Fields*, the textbook can be accessed through the Los Alamos arXiv archive. (http://xxx.lanl.gov/abs/hep-th/9912205).

The e-print archives are entirely scientist driven, and flexible enough either to co-exist with the pre-existing publication system, or to help it evolve into something better able to meet researcher needs (Ginsparg 2000).

Implications for Libraries:

Looking at the successful track record of the arXiv over the past decade, what lessons can be shared?

- Revolutions require leaders who not only have vision but also can stay the course over the long term. Continuity is critical. Paul Ginsparg and the support team he has assembled have consistently demonstrated passion and an extraordinary commitment to the scientific community.

- The Los Alamos archive is a user-created and user-driven system. It has been successful because it meets the needs of its user community. As such, it should send a strong and clear message regarding understanding user requirements when designing and maintaining digital library applications.

- From day one and still today, operational support for the arXiv requires people with very technical skills. Attracting, training and retaining people with advanced degrees in science, coding talent in Perl, Unix system administration skills, etc., is a tough challenge. Such highly talented individuals are highly sought after and worth their weight in the labor market.

- There is no substitute for good metadata. The importance of standardized metadata is not in the least surprising to librarians, but the challenge of obtaining accurate metadata from authors has been a surprise. References have to be entered correctly by the author to link correctly. Keywords have to be consistent among different articles. Others have observed that authors are not particularly good at doing these things correctly, and we have had the same experience (Boyce 2000). More automated tools can fix many things, but there is no antidote to a lack of common sense.

- Operating a worldwide international service has very different implications and requirements as compared to keeping the local system up and going on a 24 by 7 schedule.

Digital Library Intersections

Ideally, researchers desire unfettered access to the scientific literature. They have come to expect the ability to navigate via rich hyperlinks between papers published in different journals, without consideration of expensive subscriptions or the use of passwords, to access many dozens of journals in their fields. The connections with current state-of-the-art digital library services are obvious. The LANL Library Without Walls already has a very rich linking environment: transparent linking between nine large web locally mounted databases and over 4,000 electronic journals; full implementation of SFX (Van de Sompel and Hochstenbach 1999) or LinkSeeker at LANL, for context sensitive dynamic linking; and the ability for users to search across all our collections, including the arXiv via FlashPoint.

These are simply beginning, early steps in the transformation of use of the scientific literature. User adoption of these integrated capabilities has been high and corresponding user feedback has been very positive. The tighter integration of formal and newly rising informal e-print systems represents an enormous opportunity for libraries and information providers -- all to the benefit of the researcher.

Conclusion

The journey in the e-print revolution has witnessed a decade of upheaval in the scientific research community. This new technology was started and adopted by researchers frustrated with the lack of effective and efficient communication with the process of scientific publishing. As the e-print system evolved, it has become an important and fundamental source of communication for the community. It is now incorporated into our scientific research arena in a unique manner. The details change, but the underlying principles continue. As libraries evolve in the changing electronic revolution, they can continue to have an important role in supporting research.

The arXiv is based upon activities supported by the U.S. National Science Foundation under Agreement No. 9413208 (1 Mar 1995 thru 30 Sep 2000) with Los Alamos National Laboratory, and by the U.S. Department of Energy. Mirror sites are funded locally. Opinions, findings, and conclusions expressed in this article are those of the author and do not necessarily reflect the views of the U.S. National Science Foundation, the U.S. Department of Energy, or Los Alamos National Laboratory.

References:

APS forms cooperative agreement with LANL's xxx e-print archive. 1998. *APS News Online* (March 1998). [Online] Available: http://positron.aps.org/apsnews/0398/039805.html [February 21, 2001].

Boyce, Peter B. 2000. For better or worse: preprint servers are here to stay. *College and Research Libraries News* 61(5): 404-407, 414.

First online graduate physics textbook hits the web. 2000. *APS News Online* (March 2000). [Online] Available: http://positron.aps.org/apsnews/0300/030011.html [February 21, 2001].

Ginsparg, Paul. 2000. Creating a Global Knowledge Network: Don't Just Clone the Paper Methodology In: *Freedom of Information Conference:* (6-7 July 2000: New York Academy of Medicine). [Online] Available: http://www.biomedcentral.com/info/ginsparg-ed.asp [February 21, 2001].

Holtkamp, Irma S. and Berg, Donna A. 2001. The Impact of Paul Ginsparg's ePrint arXiv (Formerly Known as xxx.lanl.gov) at Los Alamos National Laboratory on Scholarly Communications and Publishing: A Selected Bibliography. [Online] Available: http://lib-www.lanl.gov/libinfo/preprintsbib.htm [February 21, 2001].

Langer, James. 2000. Physicists in the new era of electronic publishing. *Physics Today Online.* [Online] Available: http://www.aip.org/pt/vol-53/iss-8/p35.html [February 21, 2001].

Pack, Thomas and Pemberton, Jeff. 1999. A harbinger of change: the cutting edge library at the Los Alamos National Laboratory. *Online Magazine.* 23(2): 34-42. [Online]. Available: http://www.onlineinc.com/articles/onlinemag/pack993.html [February 21, 2001].

Van de Sompel, Herbert and Hochstenbach, Patrick. 1999. Reference linking in a hybrid library environment part 3: generalizing the SFX solution in the "SFX@Ghent & SFX@LANL" experiment. *D-Lib Magazine.* 5(10). [Online] Available: http://www.dlib.org/dlib/october99/van_de_sompel/10van_de_sompel.html [February 21, 2001].

Will Journal Publishers Perish? 2000. *The Economist* 355(8170): 81-82.

FEEDBACK

We welcome your comments about this article. Please fill out this form for possible inclusion in a future issue.

MAILING LIST

Would you like to be notified about new issues of ISTL? Join our mailing list.

Previous		Contents		Next

Originally published in Issues in Science and Technology Librarianship, Number 29, Winter 2001 (http://www.library.ucsb.edu/istl/01-winter/article3.html).