Building DSpace to Enhance Scholarly Communication

by Eric Celeste and Margret Branschofsky

Abstract

The MIT Libraries has built a durable digital document repository called DSpace to house the digital output of our faculty's research. DSpace will be a home for the digital documents our faculty want to share with their colleagues around the world. DSpace also expresses the ferment in scholarly communication, and the potential shift away from the journal as the primary means of disseminating research findings. This article takes a brief look back at where scholarly communication has been, describes how it may now be changing, shares our vision of how DSpace fits into that picture, and glances at the impact DSpace will have on our faculty and library.

Keywords

digital libraries, scholarly communication, digital archives, digital repositories, preprint servers

Authors

Eric Celeste, MLS, Assistant Director for Technology Planning and Administration, MIT Libraries, Bldg. 14S-308, Massachusetts Institute of Technology, Cambridge, MA 02139; efc@mit.edu

Margret Branschofsky, MSLS, DSpace Project Faculty Liaison, MIT Libraries, Bldg. 10-500, Massachusetts Institute of Technology, Cambridge, MA 02139; margretb@mit.edu

Introduction

The MIT Libraries has built a durable digital document repository called DSpace to house the digital output of our faculty's research. DSpace will be a home for the digital documents our faculty want to share with their colleagues around the world. DSpace also expresses the ferment in scholarly communication, and the potential shift away from the journal as the primary means of disseminating research findings. This article takes a brief look back at where scholarly communication has been, describes how it may now be changing, shares our vision of how DSpace fits into that picture, and glances at the impact DSpace will have on our faculty and

library.

Background

Scholarly communication has not significantly changed since the middle of the 17th century when the Royal Society of London launched its Philosophical Transactions in 1665. This development marks the beginning of the scholarly journal, whose original purpose was to formalize and regularize the exchange of information between scholars, replacing the heavy exchange of correspondence previously required to keep abreast of developments in the world of learning. Learned societies continued to be the primary publishers of scholarly journals until well into the 20th Century, providing services to members of the society and to universities, while relying on the support of scholars and universities to maintain financial security. The society journal offered scholars a vehicle for disseminating their work and ideas, while also providing them with a means of learning about others' endeavors. Authors contributed to the advancement of journals in their disciplines by providing editorial and reviewing services gratis. Development of the peer review system provided universities with an evaluation system for rewarding faculty through advancement and tenure. Universities contributed to the support of society journals through subscriptions paid by their libraries.

It was not until after World War II, when government funding of science and technology resulted in a huge increase in numbers of journals published, that this model changed. As more authors submitted articles to be published, and as disciplines became more specialized, many more journals were started. University libraries were well funded and bought these journals to keep up with demand from faculty. Such a climate of demand made it possible for commercial publishers to step into the journal market, gradually raising prices and making profits. The continuous rise in price of journals came to a head during the 1980's when university library budgets were no longer able to keep up with the inflation of prices.

Informal means of communication have always accompanied the more formal print outlets. Scholarly communication occurs at conferences, through correspondence, and through the exchange of pre-publication information within a discipline. An informal network of communication between the top scholars in a discipline, called the “invisible college”, often provides the primary means of communication between established scholars. Because of the long time lag between the submission of an article for publication in a print journal and the actual publication date, a practice of exchanging preprints became a part of the communication culture within several scientific disciplines.

As early as 1961 an early experiment to develop a centralized method of disseminating preprints via the Information Exchange Groups in the biomedical sciences was supported by the NIH for a period of six years. Also during the 1960s, the Stanford Linear Accelerator Center (SLAC) took an early lead role in collecting and cataloging preprints in the field of High Energy Physics (HEP), thereby supporting a preprint culture in that field.[1] It was in HEP that the first pre-print server was established at Los Alamos National Labs in the early 1990's.

Shift

Scientific advance supported by peer-reviewed publishing in the journal literature has come under pressure from two fronts. On one front, librarians have been concerned about the business model of journals, particularly the fact that journals have been getting too expensive for library budgets. Observers on the other front have feared that the pace of journal publishing cannot keep up with the pace of innovation. Pressures from both fronts may squeeze scholarly communication apart into its component services, ready to realign into a new system.

Herbert Van de Sompel [2] has drawn from and expanded on Roosendaal and Geurts [3] to propose six basic services which comprise scholarly communication: Registration, Awareness, Certification, Archiving, Rewarding, and Accessibility. Since these are not the most familiar terms

in this context, they deserve some definition. Registration (such as submitting a paper to a journal) is the service which takes in a scholar's work and acknowledges it as his own. Certification (such as peer-review) declares the work to be of a certain scholarly value. Others learn of the work's existence through Awareness services (such as advertising). Accessibility services (like library lending) make the work available to others. Archiving preserves the work for future scholars. And finally Rewards (such as tenure) encourage scholars to keep contributing to the system.

In the journal publishing model, Registration, Certification, and Awareness are all typically provided by publishers, Archiving and Accessibility are typically provided by libraries, and institutions usually do their own Rewarding through systems like tenure. But as many respected journals fall months and even years behind the pace of innovation in their fields, new technological approaches are pulling these services out of their comfortable contexts. Preprint servers such as arXiv.org e-Print archive (now housed at Cornell) provide Registration of new works, new forms of open Certification are emerging on the Internet, and Accessibility is easier to provide than ever, thanks to the web. But participation in this new model is limited to a few scholarly disciplines and the depth of commitment to their new roles by some of the players is uncertain. Still, new technologies offer the opportunity to redefine scholarly communication, allow faster dissemination of works, and revise business models.

As these scholarly communication services are squeezed apart and realigned, new players can take responsibility for each role. We have already seen examples like ArXiv.org emerge, dedicated to improving communications within particular disciplines. This makes a certain amount of sense, since the allegiance of faculty is often at least as great toward their discipline as it is to the institution housing them at any given moment. Still, those institutions have a great deal of interest in the output of their faculty, and may have more consistent resources available to sustain digital repositories over the long haul. Institutions also have an interest in capturing output from all disciplines, not just those with a cultural predisposition toward sharing.

While the MIT Libraries certainly do not intend to replace the journal publishing model, we do want to play a part in providing some of the component services of scholarly communication in the digital age. In particular, we don't see others systematically capturing and providing access to the digital output of our research process, output in which we have a large vested interest. As models realign we think it is vital that libraries step up to such a challenge.

Responding with DSpace

In 1999 MIT Libraries decided it was time to take some action on behalf of our institution to capture the digital output of the research that happens at MIT. The transformation of scholarly communication under way and the emergence of a community dedicated to making it possible to sew together the contents of a diverse universe of open electronic archives constituted a ripe environment for this kind of experiment. And most importantly, a corporate partner with an interest in this kind of tool came to light: Hewlett-Packard.

Hewlett-Packard Laboratories, the research arm of HP, was interested in learning what it takes to start up and manage a digital repository and how the market (in this case the MIT community) would respond to the availability of such a repository of research output. Such repositories of digital content are a key aspect of the future, and in academe HP has found a partner ready to experiment with that future today. HP Labs also wanted a testbed for further research initiatives in the capture, storage, management, and dissemination of digital documents.

Our view that it was important to attempt an institutional approach to collecting digital output of research found some quick confirmation. In the fall of 1999 the MIT Libraries received an invitation to participate in the Universal Preprint Server (UPS) meeting sponsored by a wide array of library organizations and the Research Library of the Los Alamos National Laboratory.[4] The main thrust of this meeting was the development of the "open archives protocol" and formation of

what became the Open Archives Initiative. We agreed in the now superceded Santa Fe Convention to a framework to support the basic interoperability of electronic print archives. [5] But of great interest to MIT was the tacit acknowledgment of such a wide array of participants that there were two ways to build such archives: in disciplinary stovepipes or as institutional repositories. At this point we'd already begun our negotiations with Hewlett-Packard to develop an institutional repository, and in Santa Fe we found that MIT was the only organization moving ahead with a real-world test of the feasibility of setting up an institutionally bound repository.

Defining DSpace

DSpace is intended as a home for completed works resulting from research at MIT, works ready to be shared with colleagues. DSpace will not house works-in-progress or other works not yet fit for broad exposure. DSpace will offer producers the option of superseding a work with a corrected or enhanced edition, but the older editions will remain on the system for the record. Producers shall only submit work to DSpace if their intent is to share that work with the world. Certain conditions may prevent our ability to fully share a particular work at a particular time, but such sharing should be the intent of the producer of the work.

In order to fulfill our desire to take responsibility for the Registration, Accessibility, and Archiving services of scholarly communication with respect to MIT's own digital output, we required that DSpace primarily serve as a reliable repository. In order to ensure that the services we provide could tie into those offered by other players we would also have to take advantage of existing and future protocols. To build a sustainable system would require that others adopt DSpace as a solution to similar challenges, which in turn required that we understand business models which can support such a system. These, and learning to work with a corporate partner like HP, were the essential goals with which we embarked on the project.

A reliable repository of digital works at MIT must cope with producers from a wide variety of

disciplines. Supporting these disciplines requires a variety of submission paths with differing assumptions about metadata schemas, workflow steps, and approval policies. A professor of physics will come to DSpace with very different requirements than a researcher in genetics. DSpace will have to allow them to use appropriate language to describe their works and employ separate paths to move the works out for public access.

In return for submitting their work to DSpace, faculty and researchers (our producers) will acquire a persistent URL which they can share with colleagues and cite, knowing that it will not change or become a dead link. They will benefit from preservation services which will enable future researchers to retrieve documents even if the format in which they were originally submitted is no longer supported by common tools of the day. They will find that their work becomes part of a greater body of work which will attract scholars and increase the exposure of all the material within it. They will not have to worry about maintaining their own high-availability presence on the web.

We intend for DSpace to participate in the emerging economy of open digital archives worldwide. Support for the Open Archives Protocol is a first step in this direction. As other standards emerge for the interchange of information among archives, DSpace will take advantage of them.

One indicator of DSpace success will be its adoption by other academic research institutions. The Andrew W. Mellon Foundation is supporting our adoption efforts by funding positions within the project aimed at discerning what might be appropriate business models for DSpace. In order to be able to sustain DSpace over the long haul, or to convince others that they can and should implement a DSpace of their own we must understand how we can support the long term effort of managing this repository. Part of the DSpace effort is understanding what it costs and how we'll pay for it.

Impact on Faculty

Another indicator of success will be the extent to which faculty participate in contributing content to DSpace. Faculty responses to the project have fallen into two distinct categories depending upon the publishing conventions and "cultures" within their disciplines.

In disciplines where it is established practice to submit one's papers to electronic preprint archives, such as Physics and Mathematics, faculty expressed no concerns regarding peer review, nor were they concerned about jeopardizing their chances of publishing in established journals. These faculty roundly dismissed journals that reject articles based on previous appearance on a website. Some faculty claimed that the feedback they get from dissemination on the preprint server is more valuable than comments made by the limited number of reviewers provided by publishers. They claimed that preprint servers were more important than traditional journals. On the other hand, these faculty did not see a strong need for an MIT-based system, since they already have adequate facilities for preprint dissemination through their disciplines. We were assured, nevertheless, that many of them would also contribute to DSpace, as long as we made the submission process simple. Ideally, they would like to be able to submit to their discipline-oriented e-print sites and DSpace using exactly the same procedure, preferably in one step.

In most other disciplines, where there is not a strong e-print culture, faculty expressed a strong belief in the peer-reviewed publishing process, and also expressed strong concerns about the quality of DSpace' content. Many expressed concern about copyright transfer agreements they sign with publishers that preclude or limit the right to display articles on personal or institutional websites. Others mentioned that their need for timely dissemination of research results was already being met by publishers' "preprint" sites, where accepted-but-not-yet-published articles are displayed.

Faculty in all disciplines were interested in using DSpace to display images, datasets, video and

audio files that they don't publish in established journals. Many expressed dismay at the mounting "page charges" required by publishers, especially for color graphics and images. As more and more research is being expressed in rich-media formats, DSpace offers faculty a means of publishing an unlimited amount of digital information in formats that are not usually handled by traditional publishers.

Impact on Libraries

An organization which has focused on combing the world's resources and selecting that which best serves its host institution is instead asked to share its host institution's output with the world. An institution that prides itself on proper application of metadata is instead asked to facilitate the input of metadata by novices. In some ways DSpace turns the library inside out. Yet, the mission of an academic research library is to facilitate the teaching and research of its host institution. DSpace will serve an important role in meeting this mission over the coming decades. A close look at the architecture of DSpace reveals a structure not so far removed from what we know as a library today.

Our task of selection, of sorting through the world's bounty for that most beneficial to the work of MIT, inverts with DSpace. Our selectors today work closely with faculty to understand their interests, learn what journals they consult, and anticipate the requirements of new courses they plan to teach. DSpace asks us to draw out of our faculty and researchers that which they wish to share with colleagues around the world. We must work with faculty not only to discern their needs, but also to discover what they have to offer. We become not just one of their sources, but also one of their destinations.

Today we manage a considerable budget balancing needs as expressed by our faculty and students, deploying acquisitions agents and catalogers to bring the material here to MIT and make it visible to our community. With DSpace we will be deploying some portion of our budget to

make the work of MIT visible to the world. Yesterday the assets we managed were those items we brought to MIT, today in DSpace the assets are the output of MIT itself, the fruit of its labor.

Our organizational skills, till now deployed to tame a collection of these foreign objects so that they would be useful to the institution, will have to be trained also on our own work. In fact, much of the metadata generation will be done by the producers themselves. Our task is shifting from creating the metadata itself to building systems that help others supply appropriate descriptions and classifications of information.

Still, in some ways DSpace does not differ all that much from the traditional library. We found as we applied the language of OAIS [6] to DSpace that the same mapping exercise applied to the traditional library revealed many similarities and helped clarify the roles that librarians may play as DSpace enters into production. The OAIS model does not require that the tasks it identifies be resolved electronically, so it applies quite nicely to both the digital and the physical realms of libraries.

For example, today's creators of submission information packages (catalogers) might find an appropriate role specifying the submission information package contents for DSpace. Those who manage our current catalog media and management might find a similar role managing DSpace's database. Librarians who help our community form queries that successfully retrieve information from our current systems will help consumers of the DSpace system as well. While DSpace turns us inside out in some respects, in others it will further leverage the expertise already present in the libraries.

Conclusion

The DSpace project team is currently hard at work developing the system that will fulfill these

promises. DSpace began beta-testing with a few early adopters on campus during the Spring of 2002. A full rollout at MIT should follow in the Fall of 2002. Once the MIT implementation is firmly established, we expect to start sharing DSpace software with a small select set of academic research libraries interested in adopting it as a way to capture the output of their institutions. When DSpace has been successfully implemented at one or two other research institutions, it will be available for adoption by any interested party. We are developing DSpace as an open source effort, both to reduce barriers to adoption and to increase the pool of expertise devoted to its development. Our web site at "http://www.dspace.org" provides more details about DSpace and an opportunity to sign up for periodic updates on the effort.

Much work lies ahead in the implementation and adoption of DSpace. The MIT Libraries expects plenty of surprises and challenges ahead. We look forward to the lessons we'll learn by making this effort, and to the partners we'll get to know on the journey.

Footnotes~~:~~2

1. James E. Till, "Predecessors of Preprint Servers," *Learned Publishing* 14 no. 1, January, (2001): 7-13.

2. Herbert Van de Sompel, "Preview of the Open Archives Metadata Harvesting Protocol", *CNI Fall 2000, San Antonio, Texas, December 8th 2000.*

3. Hans E. Roosendaal and Peter A. Th. M. Geurts, "Forces and Functions in Scientific Communication: an analysis of their interplay"
CRISP 97 Cooperative Research Information Systems in Physics, refereed collection of publications of invited talks of this international workshop.
eds.: M. Karttunen, K. Holmlund, E.R.Hilf, 1997, see http://www.physik.uni-oldenburg.de/conferences/crisp97/roosendaal.html (viewed May 31, 2001).

4. First Meeting of the Open Archives Initiative, Santa Fe, New Mexico, US, October 21-22 1999, UPS, 2000, see "http://www.openarchives.org/ups1-press.htm" (viewed May 31, 2001).

5. The Santa Fe Convention for the Open Archives Initiative (SFC, 2000), see

"http://www.openarchives.org/sfc/sfc_entry.htm".

6. CCSDS, 1999*. Reference Model for an Open Archival Information System (OAIS). Red Book.*

In order to describe DSpace with some efficiency, we will use the terminology of the OAIS model. This model is proving to be a reasonable common language when discussing digital archiving projects, and is helpful in this regard. However, DSpace will differ in some fairly significant architectural ways from the OAIS model, so our use of this terminology should not imply an adoption of the whole model.