DSpace: An Institutional Repository from the MIT Libraries and Hewlett Packard Laboratories

MacKenzie Smith, Associate Director for Technology

Massachusetts Institute of Technology Libraries, Cambridge, MA 02139, USA
kenzie@mit.edu

Abstract. The DSpace™ project of the MIT Libraries and the Hewlett Packard Laboratories[1] has built an institutional repository system for digital research material. This paper will describe the rationale for institutional repositories, the DSpace system, and its implementation at MIT. Also described are the plans for making DSpace open source in an effort to provide a useful test bed and a platform for future research in the areas of open scholarly communication and the long-term preservation of fragile digital research material.

1 Introduction

Scholarly communication among academic researchers has followed well-worn paths for much of the last century. Research is conducted, and its results are summarized in published monographs or in the formal, peer-reviewed journal literature of the academic field. Publication is evidence of scholarly significance and success, and journal editors and reviewers serve at the gatekeepers of the scholarly record, and therefore of the status of individual scholars. The advent of advanced computer technology, the prevalence of networking, and publishing media such as the World Wide Web have begun to erode some of these traditions of scholarly communication. First, there is the improved speed of delivery of research results on the Internet. Preprint archives in some branches of physics and mathematics have become the normal way in which scholars communicate – however most authors expect their results to appear in print journals eventually, as the published record. Second, there is a growing amount of important research that never sees formal publication – for example the white papers and technical reports of sponsored research in computer science and biotechnology. Third, the increasing availability of primary research material on the Web makes dependence on published summaries less satisfying – scholars want to verify and reproduce results, see the data underlying the research, experience the simulations and visualizations for themselves. The printed article is gradually becoming a sort of final imprimatur of the research: serving the gatekeeper function (i.e. an editor accepted the research into an exclusive journal thus verifying its importance) but not that of scholarly communication.

[1] DSpace project home page http://www.dspace.org/

M. Agosti and C. Thanos (Eds.): ECDL 2002, LNCS 2458, pp. 543-549, 2002.

The Massachusetts Institute of Technology (MIT) in Cambridge, Massachusetts is fertile ground for this phenomenon, producing some ten thousand unpublished preprints, technical reports, and white papers annually in almost every field of science, technology, business and economics. The Institute additionally produces large amounts of primary digital research material in the form of datasets (statistical, geospatial, etc.), image sets (from radars, sonars, satellites, etc.), software for simulations, visualizations, and other mechanisms of performing or disseminating research. Many MIT faculty make this material available on a departmental web site, and it represents a significant intellectual asset of the institution. In recognition of the growing importance of this material and of the role of the library in capturing and preserving it for future researchers, the DSpace project to build an institutional digital repository was created.

Scholarly communication among academic researchers has followed well-worn paths for much of the last century. Research is conducted, and its results are summarized in published monographs or in the formal, peer-reviewed journal literature of the academic field. Publication is evidence of scholarly significance and success, and journal editors and reviewers serve at the gatekeepers of the scholarly record, and therefore of the status of individual scholars. The advent of advanced computer technology, the prevalence of networking, and publishing media such as the World Wide Web have begun to erode some of these traditions of scholarly communication. First, there is the improved speed of delivery of research results on the Internet. Preprint archives in some branches of physics and mathematics have become the normal way in which scholars communicate – however most authors expect their results to appear in print journals eventually, as the published record. Second, there is a growing amount of important research that never sees formal publication – for example the white papers and technical reports of sponsored research in computer science and biotechnology. Third, the increasing availability of primary research material on the Web makes dependence on published summaries less satisfying – scholars want to verify and reproduce results, see the data underlying the research, experience the simulations and visualizations for themselves. The printed article is gradually becoming a sort of final imprimatur of the research: serving the gatekeeper function (i.e. an editor accepted the research into an exclusive journal thus verifying its importance) but not that of scholarly communication.

The Massachusetts Institute of Technology (MIT) in Cambridge, Massachusetts is fertile ground for this phenomenon, producing some ten thousand unpublished preprints, technical reports, and white papers annually in almost every field of science, technology, business and economics. The Institute additionally produces large amounts of primary digital research material in the form of datasets (statistical, geospatial, etc.), image sets (from radars, sonars, satellites, etc.), software for simulations, visualizations, and other mechanisms of performing or disseminating research. Many MIT faculty make this material available on a departmental web site, and it represents a significant intellectual asset of the institution. In recognition of the growing importance of this material and of the role of the library in capturing and preserving it for future researchers, the DSpace project to build an institutional digital repository was created.

The DSpace system is the result of a joint development project by the MIT Libraries and the Hewlett Packard Laboratories to build an institutional repository system for the research output of MIT faculty in digital formats. The system supports the capture, management, dissemination and preservation of this digital material. DSpace is about to be deployed at MIT, and we hope to extend its use to other institutions to facilitate the sharing of MIT's intellectual content and metadata and to allow the system to benefit from a larger community of users and developers. The system will ultimately be freely available under an open source license. DSpace consists of tools for the loading, administration, and dissemination of digital content following the Open Archival Information System (OAIS) reference model[2]. These tools include integrated subsystems for web-based and batch submission of digital material and related metadata, locally-configured submission workflow management, cross-system metadata schema, index and search, archival package management, access policy control, robust provenance and history logging, persistent identifiers, and administration.

The DSpace project is currently completing an early-adopter phase that has provided us with a large amount of material to both exercise and demonstrate the utility of the system. Early adopters at MIT have included the Sloan School of Management, the Department of Ocean Engineering, the Center for Technology, Policy and Industrial Development, and the Lab for Information and Decision Systems. In addition, two historical collections have been loaded: the out-of-print books of the MIT Press (in PDF format) and a large collection of technical reports from the NSF-funded NCSTRL project[3]. Each of these adopters has formed a community that defined its own membership, collections, workflow, metadata practices, and interface design. The system is scheduled for general release to the MIT community in September of 2002 and the MIT Libraries have developed a detailed transition plan for bringing the system into its daily operations. Shortly after the general MIT release, the system will be made available to some other institutions wishing to federate with the MIT Libraries and willing to help develop new functionality that a federated system might support (e.g. "virtual" distributed collections, new journal publications, cross-institutional searching, etc.). Since DSpace will ultimately be available under an open source distribution license, this could support the rapid development of both institutional digital archives and new modes of online scholarly communication

2 Information Model

The DSpace system implements an information model that will be familiar to many in the digital library field. The system is organized into "communities" which reflect the institution's organizational structure (i.e. at MIT these would be schools, departments, labs, centers, programs, etc.). DSpace communities can differ from each other in useful ways: each community defines one or more content "collections" that they would like to create (i.e. technical reports, preprints, datasets, white papers, images,

[2] Open Archival Information System http://ssdoo.gsfc.nasa.gov/nost/isoas/
[3] National Computer Science Technical Reports Library http://www.ncstrl.org/

etc.). These collections group material in whatever way the community chooses, based on their existing practices. Each community further defines a submission workflow that reflects its policies and procedures. The user roles currently defined in the workflow module include: submitters, approvers, reviewers, and editors. Communities register their members to play one or more of these roles, and define submission procedures that enforce their local policies. For example, one community might allow faculty to submit material directly to DSpace while another might define a few administrative staff to review faculty submissions, then have the department chair approve each reviewed submission. Material is ingested into DSpace only after it has passed all the workflow steps defined by the community. Finally, each community can tailor its user interface with a logo, explanatory text, and so on, while keeping within the general DSpace user interface paradigm.

Collections within communities consist of "items", each of which is a logical work (using the definition of "work" proposed by the Functional Requirements for Bibliographic Records[4] standard, as opposed to its "item" entity). Items are, in turn, composed of one or more bitstreams, or physical files of digital material. An item's bitstreams are defined by a "bundle" which will be implemented using the Metadata Encoding and Transfer Standard (METS)[5]. A couple of examples might help clarify the need for this. The simplest case of a DSpace item is a single bitstream, for example a digital image encoded as a TIFF file, or a digital document encoded as a PDF file. Another example, however, is of a PDF and a Microsoft Word version of the same work, and so both bitstreams (i.e. the PDF file and the Word file) make up the item. Another example is a print document that has been scanned into a set of digital TIFF image files, all of which together and in the correct sequence constitute the logical item. A final example is a digital document that consists of a set of several HTML pages and some inline JPG images. All of these are types of DSpace items, and there are undoubtedly other cases where physical files should be grouped for presentation to users. Digital preservation will presumably be done at the bitstream, or physical file, level in cases where format migration is done, and if the preservation strategy used is emulation then it may be done at the item level. The information model supports either approach.

For descriptive metadata, DSpace uses a qualified Dublin Core[6] vocabulary derived from the Library Application Profile for common description across all content types, and the METS framework for information packaging. Appropriate descriptive metadata about items is provided by submitters, rather than library staff, as part of the system's submission process. Submission also entails agreeing to a license that grants the library a non-exclusive right to store, preserve, and distribute the item (copyright is retained by the author or institution). Minimal technical metadata about items is captured during submission, and it's automatically generated from the physical file (i.e. the file format, size, checksum, etc.). MIT will be undertaking research in the future to determine what set of technical metadata is needed to support long-term preservation of various digital formats such as PDF, XML, etc. In addition to

[4] Functional Requirements for Bibliographic Records http://www.ifla.org/VII/s13/frbr/frbr.htm
[5] Metadata Encoding and Transmission Standard http://www.loc.gov/standards/mets/
[6] Dublin Core Metadata Initiative http://www.dublincore.org/

descriptive and technical metadata, DSpace uses the Harmony/ABC model-based mechanism for recording the history of changes within the system, which is implemented using RDF[7].

The assignment of persistent identifiers to digital items in DSpace is one of the great benefits that communities perceive. The phenomenon of "link rot" is becoming a well-known problem on the Web, especially with the sort of unpublished or semi-published material that will be DSpace's main content [1]. To avoid this, DSpace items are assigned a CNRI Handle[8] as part of the submission process, and the handle becomes part of the item's descriptive metadata. Since DSpace items can belong to more than one collection, the local handle resolver resolves requests to the item's metadata page without a collection context. From there, the user can link to the digital object itself. The handle is always displayed with the item's metadata, along with a recommendation to use the handle for citations to the item. A point of ongoing discussion has been the dual role of the handles in DSpace to serve as persistent URL and as a sort of globally unique accession number for the item. In future we may decide to assign a unique identifier to the work/item in addition to the handle to allow functions such as determining identity between multiple copies of an item that are distributed at several institutions running DSpace.

3 Architecture

Internally, DSpace can be thought of as a Digital Asset Management System that implements the OAIS reference model and includes subsystems that support common digital library functions (e.g. indexing and search of metadata, secure digital object access and delivery, collection management and preservation planning, etc.). The goal of the system is to provide MIT faculty with a robust, scalable, preservation-quality institutional repository for its born-digital research output, so that has been the initial focus of development rather than the support of digitally reformatted library collections. The system is primarily written in Java, and uses only free software libraries and tools, including the PostgreSQL RDBMS, the Lucene search engine, Xerces/Xalan XML tools, and Jena, an RDF tool from HP, among others.

The system has been implemented with clearly delineated modules, divided into user interface, business logic, and storage layers. Each module, or subsystem, has a well documented API so that they can be implemented differently at other institutions that have different practices for functions such as authenticating and authorizing users, or managing transaction logging and history tracking. We expect DSpace to grow over time, both in scope and functionality, and designed the system to be both flexible and extensible by MIT and other institutions.

In anticipation of its future deployment at other institutions through a federation, DSpace supports several of the current standards for interoperability: the Open

[7] Resource Description Framework http://www.w3.org/RDF/
[8] Corporation for National Research Initiatives Handle system http://www.handle.net/

Archives Initiative[9] protocol for metadata harvesting, and the OpenURL[10] standard among them. DSpace item metadata is available for harvesting by union catalog creators and other information service providers using OAI. MIT Libraries run the SFX[11] system from Ex Libris, and have successfully tested DSpace as both an SFX "source" and "target" to allow cross-linking between the library's online public catalog and it's DSpace institutional repository. One of the many local policies that MIT is considering is whether to catalog DSpace resources in its central library catalog or to rely on cross-system searching and linking to help users find library material in its various locations. Support will soon be added to the system for exporting items encoded in METS to support virtual collection building and other types of cross-institutional interoperation.

4 Business Plan

A business plan for the initial deployment and long-term sustainability of the DSpace system at MIT has been developed with the help of a grant from the Andrew W. Mellon Foundation. This will allow MIT to decide how to fund DSpace as a service of the Libraries using detailed cost information from MIT's own implementation, and will potentially help other institutions determine what the costs of running an institutional repository in their own organization might be. Cost and business models are one of the aspects of digital library research and development that has gotten the least attention, yet in the long run will be critically important to making these systems sustainable and justifying the large investments required in their development. We hope to share the information gathered in the MIT business plan, and to build on it for further research projects that investigate the actual costs of managing digital collections and performing digital preservation over the long-term. We will also document the results of our service model (including both free and for-fee services) to help the community understand where value-added services may be created to help pay for the operation of institutional repository systems.

5 Conclusion

In conclusion, MIT Libraries and the HP Laboratories have developed a new system that will serve as an institutional repository for the MIT faculty's intellectual output in digital formats. This repository will be the basis for managing the institution's digital research collections over time and for preserving this fragile material. DSpace will be tested at MIT over the coming year and will be made available to other institutions wishing to archive digital materials and possibly federate their collections. DSpace will be an important new platform in the coming years for several areas of digital library research including digital archiving and preservation, and new modes of scholarly communication.

---

9 Open Archives Initiative http://www.openarchives.org/

10 OpenURL http://library.caltech.edu/openurl/

11 SFX http://www.sfxit.com/

References

1. Kiernan, Vincent. Nebraska Researchers Measure the Extent of ‘Link Rot’ in Distance Education. Chronicle of Higher Education [Internet]. 2002 Apr 10 [cited 2002 May 4]: Available from http://chronicle.com/free/2002/04/2002041001u.htm