

DSpace for E-Print Archives

MacKenzie Smith (*)

Abstract:

DSpace™ (<http://dspace.org/>) is the new open source digital repository system from the MIT Libraries and Hewlett-Packard Labs designed to support the digital collections of academic research institutions, as well as the SPARC conception of Institutional Repositories for digital research material [1]. The DSpace system has been described elsewhere in detail [2] so the focus of this article is on its implementation at MIT for archiving e-prints and other artifacts of scholarly communication, and making these available to the public. The MIT Libraries are deeply concerned about the well-documented crisis in scholarly communication [3] and are committed to working towards innovative solutions. We share this concern with many of the MIT faculty and administration, several of who have been key supporters of the DSpace project and related initiatives at the university. The MIT Libraries were a founding member of SPARC, and are a signatory of the Budapest Open Access Initiative (BOAI). This article will describe how MIT Libraries have implemented DSpace to support these goals.

Introduction

DSpace™ is an open source system developed by Hewlett-Packard Labs and the MIT Libraries and available at <http://sourceforge.net/projects/dspace/> which was designed for use by academic research institutions that wish to capture, archive, preserve, and make available the scholarly research material produced by their faculty and researchers. The system itself is a simple, but fully-featured, digital asset management system, including a submission system that supports complex, flexible workflows, as well as limited support for access control and delivering complex digital content. DSpace can serve a variety of types of organizations to manage their digital assets, but it was designed and optimized for academic research institutions to manage their digital research materials.

DSpace is being used to develop a range of new services within the domain of research institutions. Among other services, research libraries are using DSpace to host digital research data (e.g. images and datasets), electronic records, digital library collections, and teaching material. But one of the dominant uses of the platform is to host digital documents, be they unpublished grey literature or published research articles. Much of this material is being collected under the model of faculty "self-archiving" where faculty authors retain copyright to their published articles, or at least the right to make an electronic copy available from a website at their own institution free of charge. This use of DSpace in support of the Open Access movement is one that has the potential to transform scholarly communication in the future (making access to research results easier, faster, and cheaper) and to finally begin to change the current dynamics of the "journal crisis".

Institutional Repositories

The SPARC organization has defined Institutional Repositories to have the following properties [4]:

- They are bounded by an institution (i.e. the content is generated by the institutional community)
- They contain scholarly content (e.g. preprints and working papers, published articles, enduring teaching materials, student theses, datasets, etc.)
- They are cumulative and perpetual (i.e. they preserve ongoing access to material)
- They are interoperable and open access (access is no- or low-barrier, online, global)

The DSpace system was designed to support each of these properties, although institutions aren't required to use it in this way. The technology itself is neutral, and we encourage organizations to explore what will work best for their own research community and communication needs. For academic research institutions wishing to make their scholarly research material more widely and easily available, DSpace offers a ready-made solution.

Even before choosing a technology platform to host the Institutional Repository, libraries or other organizations within an institution must carefully consider their service model, policies, and business plan for the Repository. Running an Institutional Repository, even one like DSpace that is free and open source, takes resources and careful planning to implement and operationalize. Consulting the SPARC Institutional Repository Checklist & Resource Guide [5] and the DSpace@MIT business plan [6] among other documents can help organizations develop local policy and think about how to make the system sustainable over archival time frames (i.e. hundreds of years).

Institutional repositories are a new and evolving concept, and we expect to see dramatic changes over the coming decade as our understanding of the issues involved with these services grows from practical experience. This article addresses only a snapshot of what will be a long conversation among librarians, archivists, authors, publishers, and technologists, and policy makers.

The DSpace Information Model

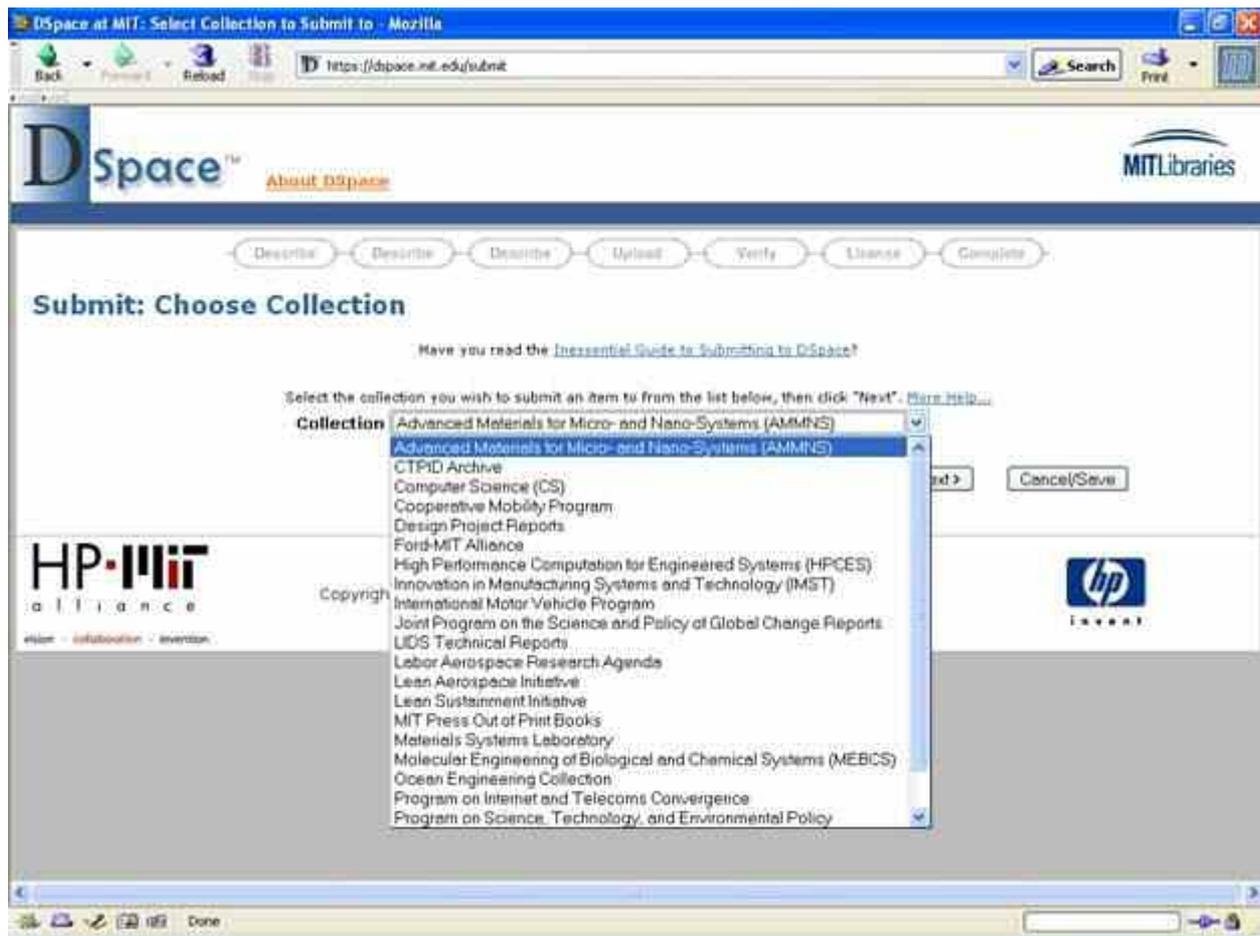
The DSpace system is designed to model the structure of an organization: it includes one or more "Communities" which are mapped to parts of the organization. In MIT's case these are the schools, academic departments, research labs and centers of the Institute. Each of these communities sets many of its own policies to reflect the culture of the particular academic discipline. They define what "Collections" will be included in their community, be they e-prints, theses, datasets, working paper series, technical reports, visualizations for teaching, etc. Communities also define two important things: who from the community can contribute items to each of its collections, with what workflow to process and approve them, and who will have access to those collections: the general public, the MIT community, or only the collection's administrators. Within a collection there are "Items" which represent the digital documents that users interact with (e.g. papers, datasets, images, visualizations, websites, and so on). Items are often associated with just one digital file or "bitstream", but can sometimes be comprised of multiple bitstreams (e.g. a web site with several HTML pages and some JPG images, or a scanned document with a TIFF image for each page). There are also cases where an item has multiple versions available, whether because different physical versions are available (e.g. PDF, Microsoft Word, and TeX) or because different versions

of a document have been released over time (e.g. an preprint and subsequent e-print version of an article). DSpace tries to accommodate all these variations of practice so that the same system will work for widely varying academic disciplines and cultures, from particle physicists to ethnomusicologists to theoretical economists.



<https://dspace.mit.edu/handle/1721.1/721>

Depositors for each item submitted also provide metadata about that item. DSpace currently uses a set of descriptive attributes drawn from a well-known schema called the "Dublin Core" which consists of familiar information such as title, author, creation date, subject, format, and so on. At MIT it is depositors who supply this information, but other institutions are choosing to have the library staff do this work on behalf of the faculty, to lower the barrier of use even further. A third scenario involves using existing metadata and converting it to the required elements for DSpace. In practice, this has often been possible at MIT, drawing from the Libraries' online public catalog or from existing databases or Websites within the Community. When depositors do have responsibility for providing metadata the quality can vary widely, although we are seeing higher-than-expected quality and completeness of the provided metadata. This is probably due to the Community's desire to have their materials be more publicly accessible which they understand to be a function of the amount and quality of the metadata they provide to the system.



Collection choice for document submission

DSpace currently has no special tools for data rendering: whatever was deposited is made available to users over Web browsers exactly as it was supplied. Fortunately, most document formats in use today are "Web native" so they can be rendered directly by common Web browsers without the need for custom software to view them. As long as deposits to DSpace continue to be in formats like Adobe PDF or HTML there will be little problem with reading them.

MIT and Open Access

MIT as an institution is committed to open access to research and teaching information as a core part of its mission. In addition to DSpace there are several high-profile initiatives underway at MIT to increase access to the Institute's many scholarly resources. Of the OpenCourseWare project which seeks to make all of MIT's course material available on the web for free, MIT's president Vest says "MIT OpenCourseWare reflects the commitment of the MIT faculty to advancing education by increasing access to their academic materials through the Internet and the World Wide Web. We believe that with modern communication technology we can not only transmit information but also stimulate and enhance the deeply human, person-to-person endeavor of education. We hope the idea of openly sharing course materials will propagate throughout many institutions and create a global web of knowledge that will enhance the quality of learning and, therefore, the quality of life worldwide."

This philosophy is shared by many at MIT, and also directly affects the goals of DSpace. President Vest also said: "Today's stewardship of accessible knowledge is inherently interdisciplinary and necessarily connects the full range of activities from archiving to publishing. University research librarians," Vest concluded, "are central to managing this complex range of activities and can play a major role in accelerating efforts toward the open sharing of knowledge." [7]

In an effort to forward these goals, the MIT Libraries' initial policies for the DSpace service favor this aspect of use: deposits are limited to scholarly works and teaching material authored by MIT faculty and researchers (not students, except for theses, and not administrative records). The service is also designed to make the barriers to use by faculty as low as possible: DSpace accepts any technical format of material, not just those that will be straightforward to migrate for continued access over time (e.g. proprietary formats like Microsoft Word are accepted).

Faculty Reaction

After almost a year of running DSpace as a production service of the MIT Libraries, we have found that to gain acceptance by faculty to submit articles into an institutional repository it is important to offer a service in which faculty find immediate value. For example, there is little incentive for faculty to deposit their published articles today, but great incentive to deposit material they perceive as being more at risk or difficult to manage, particularly teaching material and research data. We believe that providing services to help faculty with their current problems (i.e. managing, distributing and preserving research and teaching materials) increases their likelihood of contributing articles as well, as they become accustomed to using the archive [8].

We are also discovering a wealth of important research documents that are relatively hidden at the institution now: the so-called grey literature. These are documents like working papers, technical reports, conference proceedings, and the like which are not formally published, and so not consistently acquired by the library, but are of increasing importance to scholarly communication in a wide range of academic subjects. DSpace is proving to be a valuable tool to capture and manage this type of material, thus solving a problem for the departments, research labs and centers that produce them, and making the research work of the institution more generally visible. This is perceived as a straightforward way to promote the research of a given community and to broaden its impact.

Finally, faculty are aware of the crisis in scholarly communication to varying degrees, and this educational process will take time. Faculty vary widely in their attitude towards open access to their research output (papers, but also books). Having the system support limited access controls to collections when the authors request it is necessary to reassure them that their material is being treated as they feel appropriate. Institutions that attempt to force faculty to "give away" their research articles may find considerable resistance from many faculty at this time. It is our hope and expectation that as faculty become more accustomed to seeing research articles and other works of scholarly freely available on the Web it will be straightforward to remove these access barriers - without the barriers it would be difficult to get some faculty material at all. Given our understanding of the fragility of digital documents, and our goal of preserving as much as possible of the scholarly record, it is important to capture these documents at the point of publication even if they can't be made freely available immediately.

The contents of the DSpace repository at MIT have been on a slow but steady growth curve since the system went live last year. We are seeing steady growth in deposits from the existing "early

adopter" Communities, as well as a handful of new Communities joining the system, and about a dozen further Communities in discussion about joining at any given time. The process of joining is still fairly slow, partly as a result of the expected limits on people's time to work with us, and partly a factor of our gaining experience with the process and the policies we need to make joining as straightforward as possible.

Conclusion

The ambition of DSpace at MIT, and at a growing number of other institutions, to make our research material in many forms publicly available is proving to be quite achievable. However finding incentives for faculty to participate, and working with publishers and government regulators to insure that these efforts continue to be allowed and, hopefully, become easier, is a long process that has only just started. But having a platform available with which to experiment and build these services at research institutions is going to support unprecedented progress in that process. There is nothing so rewarding as being able to get started and see what happens. We look forward to getting more institutions involved, and to working together with them to achieve real, and long-overdue, change.

To find out more about the DSpace software and how it's being adopted at a number of research institutions, visit <http://dspace.org/>. The software is freely available at <http://sourceforge.net/projects/dspace/>.

References

- [1] <http://libraries.mit.edu/dspace-mit/what/definition.html>
- [2] See, for example, "DSpace: An Open Source Dynamic Digital Repository", MacKenzie Smith et al., D-Lib Magazine, January 2003. <http://www.dlib.org/dlib/january03/smith/01smith.html>
- [3] "Scholars Under Siege: The Scholarly Communication Crisis" <http://www.createchange.org/faculty/issues/quick.html>
- [4] <http://www.arl.org/sparc/IR/ir.html>
- [5] http://www.arl.org/sparc/IR/IR_Guide.html
- [6] <http://libraries.mit.edu/dspace-mit/mit/plan.html>
- [7] <http://www.arl.org/newsltr/224/activities.html>
- [8] Clifford A. Lynch, "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age" ARL: A Bimonthly Report on Research Library Issues and Actions from ARL, CNI, and SPARC 226 (February 2003): 1-7 <http://www.arl.org/newsltr/226/ir.html>; Raym Crow, "The Case for Institutional Repositories: A SPARC Position Paper." Washington: Scholarly Publishing & Academic Resources Coalition, 2002.

Author Details

MacKenzie Smith

Associate Director for Technology, MIT Libraries: <http://libraries.mit.edu/>

Email: <http://library.cern.ch/HEPLW/9/papers/3/kenzie@mit.edu>

Address: MIT Libraries, 77 Massachusetts Avenue, Building 14S-308. Cambridge, MA 02139

MacKenzie Smith manages technology for the MIT Libraries and is currently acting as the DSpace project director. She oversees both MIT's use of DSpace and its transition to an open source, jointly maintained system used by institutions throughout the world. She also manages the digital library research program at MIT which is building on the DSpace platform to explore issues such as metadata interoperability, digital preservation, and alternative publishing models.

For citation purposes:

Mackenzie Smith "*DSpace for E-Print Archives*", High Energy Physics Libraries Webzine, issue 9, March 2004. URL: <http://library.cern.ch/HEPLW/9/papers/3/>