D-Lib Magazine
January 2003

Volume 9 Number 1

ISSN 1082-9873

DSpace

An Open Source Dynamic Digital Repository

MacKenzie Smith
Associate Director for Technology
MIT Libraries
<kenzie@mit.edu>

Mary Barton
Senior Business Strategist
MIT Libraries
<mbarton@mit.edu>

Mick Bass
HP External Engagement Manager
Hewlett-Packard Labs
<mick.bass@hp.com>

Margret Branschofsky
DSpace User Support Manager
MIT Libraries
<margretb@mit.edu>

Greg McClellan
DSpace Systems Manager
MIT Libraries
<gam@mit.edu>

Dave Stuve
Senior Developer
Hewlett-Packard Labs
<david.stuve@hp.com>

Robert Tansley
Lead Developer
Hewlett-Packard Labs
<robert.tansley@hp.com>

Julie Harford Walker
Senior Business Strategist
MIT Libraries
<jharford@mit.edu>

Abstract

For the past two years the Massachusetts Institute of Technology (MIT) Libraries and Hewlett-Packard Labs have been collaborating on the development of an open source system called DSpace™ that functions as a repository for the digital research and educational material produced by members of a research university or organization. Running such an institutionally-based, multidisciplinary repository is increasingly seen as a natural role for the libraries and archives of research and teaching organizations. As their constituents produce increasing amounts of original material in digital formats—much of which is never published by traditional means—the repository becomes vital to protect the significant assets of the institution and its faculty. The first part of this article describes the DSpace system including its functionality and design, and its approach to various problems in digital library and archives design. The second part discusses the implementation of DSpace at MIT, plans for federating the system, and issues of sustainability.

DSpace Definition, Features and Functionality

In March 2000, Hewlett-Packard Company (HP) awarded $1.8 million to the MIT Libraries for an 18-month collaboration to build DSpace™, a dynamic repository for the intellectual output in digital formats of multi-disciplinary research organizations. HP Labs and MIT Libraries released the system worldwide on November 4, 2002, under the terms of the BSD open source license [1], one month after its introduction as a new service of the MIT Libraries. As an open source system, DSpace is now freely available to other institutions to run as-is, or to modify and extend as they require to meet local needs. From the outset, HP and MIT designed the system to be run by institutions other than MIT, and to support federation among its adopters, in both the technical and the social sense. The DSpace Federation will be explored in a later section.

So what is DSpace? It is an attempt to address a problem that MIT faculty have been expressing to the Libraries for the past few years. As faculty and other researchers develop research materials and scholarly publications in increasingly complex digital formats, there is a need to collect, preserve, index and distribute them: a time-consuming and expensive chore for individual faculty and their departments, labs, and centers to manage themselves. The DSpace system provides a way to manage these research materials and publications in a professionally maintained repository to give them greater visibility and accessibility over time.

DSpace was built breadth-first: it supports every function that a research organization needs to run a production digital repository service, but as simply as possible. The project focus was on building a production quality system. It complements and was influenced by previous research in computer science and digital library architectures [2]. Our goals were to build a system that: would be immediately useful at MIT, and hopefully at other institutions; could be expanded and improved over time; and could serve as a platform for future research. With the help of developers at other institutions that adopt DSpace under its open source license, we will work to add features and improve the different functions of the system as we learn what users actually want, and how to best support such complex requirements as digital preservation and digital rights management.

DSpace is designed to make participation by depositors easy. The system's information model is built around the idea of organizational "Communities"—natural sub-units of an institution that have distinctive information management needs. In the case of MIT (a large research university) "Communities" are defined to be the schools, departments, labs, and centers of the Institute. Each Community can adapt the system to meet its particular needs and manage the submission process itself.

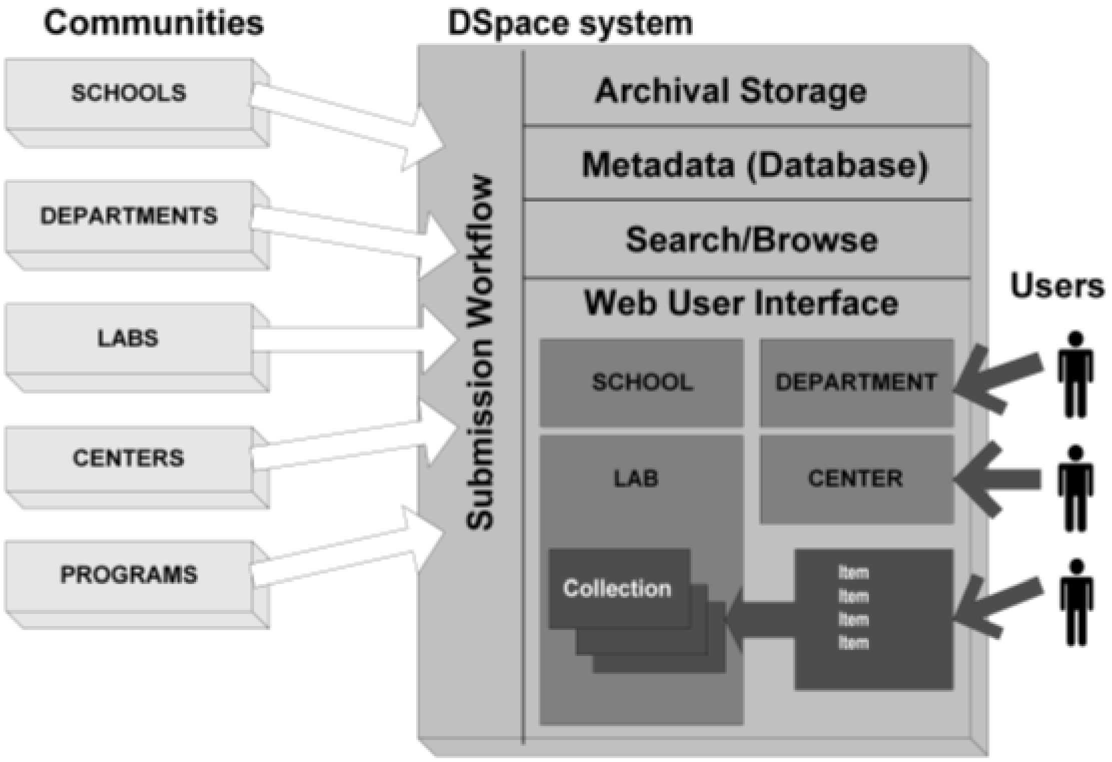

Figure 1: DSpace information model

Metadata

DSpace uses a qualified Dublin Core metadata standard for describing items intellectually (specifically, the Libraries Working Group Application Profile). Only three fields are required: title, language, and submission date, all other fields are optional. There are additional fields for document abstracts, keywords, technical metadata and rights metadata, among others. This metadata is displayed in the item record in DSpace, and is indexed for browsing and searching the system (within a collection, across collections, or across Communities). For the Dissemination Information Packages (DIPs) of the OAIS framework, the system currently exports metadata and digital material in a custom XML schema while we work with the METS [3] community to develop the necessary extension schemas for the technical and rights metadata about arbitrary digital formats.

User Interface

DSpace's current user interface is web-based. There are several interfaces: one for submitters and others involved in the submission process, one for end-users looking for information, and one for system administrators.

The end-user or public interface supports search and retrieval of items by browsing or searching the metadata (all fields for now, and specific fields in the near future). Once an item is located in the system, retrieval is accomplished by clicking a link that causes the archived material to be downloaded to the user's web browser. "Web-native" formats (those which will display directly in a web browser or with a plug-in) can be viewed immediately; others must be saved to the user's local computer and viewed with a separate program that can interpret the file (e.g., a Microsoft Excel spreadsheet, an SAS dataset, or a CAD/CAM file).

Workflow

DSpace is the first open source digital repository system to tackle the complex problem of how to accommodate the differing submission workflows needed for a multidisciplinary system. In other words, different DSpace Communities, representing different schools, departments, research labs and centers, have very different ideas of how material should be submitted to DSpace, by whom, and with what restrictions. Who is allowed to deposit items? What type of items will they deposit? Who else needs to review, enhance, or approve the submission? To what collections can they deposit material? Who can see the items once deposited? All of these issues are addressed by the Community representatives, working together with the Libraries' DSpace user support staff, and are then modeled in a workflow for each collection to enforce their decisions. The system models "e-people" who have "roles" in the workflow of a particular Community in the context of a given collection. Individuals from the Community are registered with DSpace, then assigned to appropriate roles.

For example, a department may choose to have two collections: one for working papers and another for datasets. They may then decide that any member of the faculty can deposit items to either collection directly, and that any member of the general public can have access to these collections. In this example the workflow is very simple, and the only "role" is that of submitter.

In a more complex example, the same department may have a working paper collection that requires tight editorial control by the head of the department. In this case, they may choose to again designate all faculty as "submitters", but also designate a small group of people as "reviewers", an administrative staff person as a "metadata editor", and the head of the department as the final "coordinator". An item deposited by a faculty member would then go through a process of review, cleanup and approval before finally being deposited to the relevant DSpace collection. Each person with a role to play in this process is notified of the new submission, and goes to a personal workspace in the system to perform their assigned task. Items that do not make it through the process are not archived in the system.

Technology platform

DSpace was developed to be open source, and in such a way that institutions and organizations with minimal resources could run it. The system is designed to run on the UNIX platform, and comprises other open source middleware and tools, and programs written by the DSpace team. All original code is in the Java programming language. Other pieces of the technology stack include a relational database management system (PostgreSQL), a Web server and Java servlet engine (Apache and Tomcat, both from the Apache Foundation), Jena (an RDF toolkit from HP Labs), OAICat from OCLC, and several other useful libraries. All leveraged components and libraries are also open source software. Libraries are bundled where possible (exceptions are described in the installation instructions). The system is available on SourceForge [4], linked from both the DSpace informational web site [5] and the HP Labs site [6].

While DSpace is open source and freely available, neither MIT Libraries nor HP offer formal support for DSpace adopters. It is our assumption that institutions that use DSpace will have resources to use the system, including adequate hardware that runs the UNIX operating system, and a UNIX systems administrator to install and configure the system [7]. Most institutions using DSpace will also want the services of a Java programmer who can localize and customize for them, or enhance it, although this is not absolutely necessary to run the system.

As DSpace continues to be improved by staff at HP, the MIT Libraries, and other institutions that adopt it during the coming year, MIT will take responsibility for evaluating and reincorporating these improvements into the main open source system available to the public. Plans for building a more sustainable open source maintenance strategy through the DSpace Federation will be discussed later.

System Architecture

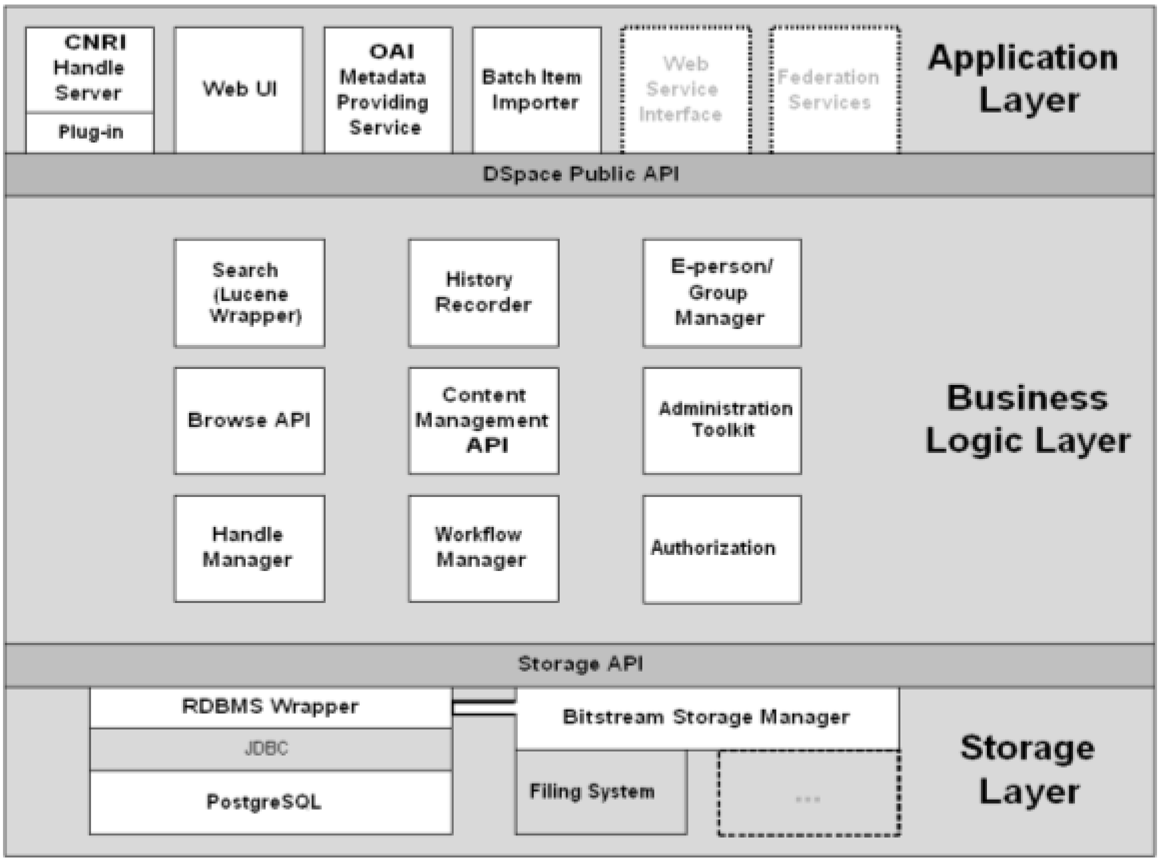

Figure 2: DSpace technical architecture

The DSpace architecture is a straightforward three-layer architecture, including storage, business, and application layers, each with a documented API to allow for future customization and enhancement. The storage layer is implemented using the file system, as managed by PostgreSQL database tables. The business layer is where the DSpace-specific functionality resides, including the workflow, content management, administration, and search and browse modules. Each module has an API to allow DSpace adopters to replace or enhance that function as desired. Finally, the application layer covers the interfaces to the system: the web UI and batch loader, in particular, but also the OAI support and Handle server for resolving persistent identifiers to DSpace items. This is the layer that will get much of the attention in future releases, as we add web services for new features (e.g., to support interoperation with other systems) and define Federation services across the range of institutions adopting DSpace.

Open Archives Initiative (OAI)

To further its goal of supporting interoperability with other DSpace adopters, and with other digital repositories, preprint, and e-print servers, the system has implemented the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [8]. DSpace used the OCLC OAICat [9] to accomplish this, and is currently exposing Dublin Core metadata for every item in the system. For material that is restricted to local access, the item metadata is exposed to OAI harvesters but the system will enforce the restriction when a user requests the associated bitstream(s). DSpace at MIT has recently been added to the OAI registry, and as the system is deployed at other institutions, we intend to investigate what added-value services might be built on top of this promising piece of infrastructure to work across the Federation. For example, we may examine the possibility of defining and building preprint and e-print collections for a particular academic discipline with individual items distributed among many institutionally-based multidisciplinary repositories, all OAI compliant.

Persistent Identifiers (Handles)

One goal of persistent digital repositories is that it be possible to find and retrieve deposited items far into the future. In particular, it is considered crucial that citations to archived material, whether found in printed articles or online, remain valid for long periods. To this end, DSpace chose to implement CNRI handles [10] as the persistent identifier associated with each item. The Handle System® covers assignment, management, and resolution of these persistent identifiers (or "handles"). Although CNRI has not registered with the IETF for an official namespace, handles are compliant with the IETF's Uniform Resource Name (URN) specification.

Handle resolution can be done using a special client, or handles can be packaged in the form of URLs and a proxy server used to resolve these into the handle form, which is, in turn, resolved to the local system location for the item. This second approach is the one we have taken in DSpace. The main alternative to using handles is to use persistent URLs with HTTP redirection to allow items to move around over time. The long-term viability of these alternatives is not yet sufficiently understood.

We plan to discuss this decision and its implications with other institutions adopting DSpace over the coming year, to see if the DSpace Federation can support other systems of persistent identification while supporting distributed services.

MIT Libraries' DSpace Implementation

DSpace is a system, a tool, and a platform for collecting, managing, indexing, and distributing digital items. Exactly how it is used, for what sort of digital material, by whom, for how long, and so on, are policy issues to be decided by each organization adopting the system. In order to make the difference between system and policy more transparent, and to help other institutions get started, MIT is openly sharing its own policy decisions with regard to DSpace. It is our hope that, while we acknowledge that our policies may not work well for other institutions, and will certainly evolve over time, they may offer guidance to others regarding the depth and breadth of issues that should be considered.

Collections Scope

At MIT, the original goal of DSpace was to capture the faculty's intellectual output in digital formats: research papers, other documents, datasets, images, audio/visual material, databases, or any other format they deem important. This goal led to two important policies: only *faculty* research would be accepted (not student material, not institutional records, and not material from non-faculty researchers without sponsorship from faculty), and *faculty* would choose what would be submitted (within certain general constraints set by the Libraries and Archives).

As a result of discussions with faculty, early adopter Communities, and others, the goal is unchanged but the policies have evolved. The first change was in what could be submitted. If a DSpace Community defines a collection that, in order to be useful, should include material authored by non-faculty (or non-MIT faculty) then it can be deposited by that Community as long as the necessary copyright permissions are obtained. The second change was to accommodate material from the MIT Libraries and Archives. We will create a Libraries and Archives Community to hold digital collections of material such as e-theses and reformatted images—material that is heavily used and represents valuable assets of the institution.

Beyond faculty-authored documents and data, another category of material has taken the spotlight for possible support by DSpace: educational material, or "Learning Objects". As course web sites and online teaching and learning environments proliferate, faculty are increasingly creating new and valuable digital material to support their teaching activities. These can take the form of traditional lecture notes, sample exams, and course calendars, but also include things like complex simulations and visualizations, multimedia presentations, or videos of key lectures. As a matter of local policy, the MIT Libraries will accept this type of material and is actively collaborating with two MIT-based projects in this area: the Open Knowledge Initiative (OKI) [11] and OpenCourseWare (OCW) [12]. For OKI, DSpace could serve as an active repository of course "content items"—those items of persistent, ongoing value (e.g., a physics simulation used regularly in various courses). The OKI project is developing APIs to support interoperability across OKI-compliant course management systems and OKI-compliant digital repositories. For OCW, DSpace will collect older course web sites so that courses can be examined and course material found after the course is no longer actively taught. Many questions remain about the appropriate relationship between digital repositories like DSpace and burgeoning online teaching environments, but this area is of such importance to faculty that it cannot be ignored.

Faculty engagement

There are several ways to describe the value of an institutional repository to the faculty who will contribute material, and the administration that will support the effort. And it is critical to explain those benefits, and to market the service, to both constituencies.

As a multidisciplinary repository that represents the scholarship of MIT, DSpace at MIT showcases the international prominence of our faculty both individually and collectively. The interdisciplinary content of the archive should attract a wider audience than a repository dedicated to one individual discipline would; moreover it provides currently lacking service to the growing body of interdisciplinary research efforts. The ability to distribute research results quickly will emphasize the cutting-edge nature of MIT's research, and supports the mission of the Institute to generate, disseminate, and preserve knowledge [13].

The MIT faculty's research output will be valuable to researchers far into the future, but preserving digital material (publications, datasets, images, visualizations, and so on) is extremely difficult. To ensure long-term access to this important scholarship the MIT Libraries will manage DSpace as a preservation archive, keeping this material accessible, and often immediately usable, far into the future.

The Libraries provide guidance in establishing new Communities, and assistance to faculty and others in using the system. DSpace was envisioned by the MIT Libraries as a continuation of their mission to collect, make available, and preserve important scholarly material of all kinds, especially that of MIT's own faculty and research community. The Libraries are working to extend their services in the digital era, to reflect current trends in scholarly communication and education, and to offer new means of distributing research material that are enabled by network technology.

Over the past few years MIT has been placing new emphasis on educational technology with initiatives such as OpenCourseWare and Open Knowledge Initiative. Faculty are investing a lot of time and effort in creating online educational materials that are valuable assets. DSpace is collaborating with the major educational technology initiatives at the Institute, including OpenCourseWare, so that storing, relocating, reusing and repurposing course content becomes reliable and easy.

Faculty accustomed to finding documents online, whether published or pre-publication, expect to continue to work with discipline-defined collections. DSpace can store and deliver preprints and eprints from the host institution and could support virtual collections from different academic disciplines by means of federation across large numbers of participating institutions. Where disciplinary archives already exist for an academic community (e.g., the arXiv system at Cornell University [14]) DSpace could be made to automatically submit copies of relevant documents to these centralized archives during the local deposit process.

Transition Team and Business plan

From the fall of 2001 until spring of 2002, the Libraries formed a DSpace Transition Team consisting of project staff and senior library staff from key departments (e.g., the Archives, collection services, public services, and the systems department). This group was charged with figuring out how to deploy DSpace as a new service of the MIT Libraries: the necessary policies, staffing requirements, communications strategies, management and governance structures, training plans, and operational requirements. Participation in this group proved to be a useful vehicle for the library staff to become more familiar with the system, and discussions of these various issues were invaluable to the development of the production DSpace service.

Participating in the Transition Team group were two senior business consultants funded by a grant from the Andrew W. Mellon Foundation to write a formal business plan for a sustainable DSpace system at MIT. Their work consisted of compiling the results of the transition team deliberations and decisions, incorporating the work into detailed cost information for system operation, and outlining possible revenue options.

The major conclusion of this planning process was that DSpace at MIT would be offered as a combination of subsidized core services (built into the Libraries' operating budget), and cost-recovered premium services that would allow the Libraries to meet varying unique needs for DSpace from particular Communities (e.g., exceptional amounts of disk storage, assistance with metadata creation, or conversion of files to supported formats). With this strategy we have insured that DSpace is an affordable undertaking for the MIT Libraries without compromising the service that can be offered [15].

Preservation

Recent discussions of digital preservation focus on at least two levels: "bit preservation", where a digital file is carefully preserved exactly as it was created without the slightest change, and what we'll refer to as "functional preservation", where the digital file is kept useable as technology formats, media, and paradigms evolve. In the first case, it's very unlikely that the file could still be read or processed by software after five or ten years have passed, but we assume it's possible for "digital archeologists" to work with the file to try to unlock its secrets many years later, especially if they have some additional information about the format (e.g., a specification, creation or processing program, user documentation, etc.). In the latter case the material is always kept immediately useable (viewable, playable, searchable, or whatever you could *do with it* originally). Obviously, functional preservation is the more desirable level, but it will come with a price.

As a community, our understanding of functional digital preservation is at an interesting juncture: we know how important the need is, we know how it can be done at an abstract level (e.g., format migrations or complex system emulation and so on). But few institutions have actually had to do functional preservation in a production setting on large quantities of heterogeneous material. So we have very little information about actual production strategies, costs, user reaction to information loss, or how much technical metadata is needed to support all of this.

How does this all relate to DSpace? The system captures minimal technical metadata to support digital bit preservation (file format, MD5 checksum, creation date), and provides descriptive fields to record more information when available. With this metadata and proper production procedures (e.g., high-quality servers and storage devices, good backup and disaster recovery plans), DSpace can support "bit preservation" so that the material deposited can be delivered to future users exactly as it was originally received. For some digital formats this may be the best option available—for example, an executable program for which no corresponding source code was provided or a format that's so rare (or proprietary) that the DSpace host institution has no way of knowing how to provide functional preservation.

However, functional preservation is currently a matter of institutional policy, and will only be implemented more thoroughly in DSpace when we understand more about the production techniques, user requirements, and cost/benefit tradeoffs. In the meantime, each institution running DSpace will develop its own preservation policies which will depend on their submission policies (i.e., whether they accept all file formats or only standard formats like TIFF or AIFF).

MIT plans to provide functional preservation for a list of "supported" formats, listed on the web site and shown to users during the deposit process. Supported formats include those that are documented standards (e.g., TIFF, AIFF, XML) or have published specifications (e.g., PDF, RIFF). The other two categories of support for MIT's DSpace are "known" and "unsupported". "Known" formats are those that are common enough to be familiar and usually quite popular, but which are proprietary in that there are no published specifications on which to base functional preservation. "Unsupported" formats are those that are either unknown to the Libraries or are extremely rare (e.g., a compiled program, a commercial CAD/CAM file, etc.). The reason for distinguishing between "known" and "unsupported" is that for the former we expect to see commercial conversion programs become available as these formats become obsolete since there are so many files in these formats in existence with many industries dependent on them. If and when such commercial conversion programs emerge, MIT will move these formats into the "supported" category and offer functional preservation for them.

The DSpace Federation

Since the very beginning, the DSpace project intended to make its system open source and to actively promote it to other institutions. Why? There are many reasons for taking this approach:

- Developing a critical corpus of content that represents the intellectual output of the world's leading research universities
- Promoting the continued development of the DSpace service through the open source community
- Promoting interoperability of archival repositories and long-term preservation of scholarly work

In 2002, MIT formed collaborative partnerships with a small number of other academic research institutions in the US, UK, and Canada, to address some specific questions such as: what will it take to successfully deploy the system at another institution? How much localization, how much customization, and how much time and effort are needed? What services can be defined to leverage the digital collections of these institutions, and how can they be implemented in DSpace? What sort of organization will the Federation become: A consortium? A new membership organization? An informal and loose collaboration? Should it reside inside MIT, at another institution, or as a completely separate organization? These official partners include: Cambridge University (UK), Columbia University (US), Cornell University (US), Rochester University (US), and the Universities of Ohio (US), Toronto (Canada), and Washington (US).

In addition to these formal collaborations, many organizations have downloaded the DSpace system (almost 1,500 since early November) and many of these are in the process of evaluating it for adaptability to their local requirements. Clearly there is great need for a system like DSpace, and as we explore the definition of the DSpace Federation over the coming year, we hope to get feedback and advice from many of these institutions about how the system should evolve and how to make it sustainable beyond MIT.

Conclusion

Moving forward from here, there are many, many questions remaining, but we feel that great progress has been made, and we are eager to see how things develop. At MIT we are very pleased and excited to have a platform to begin exploring these issues, both within the Institute and with other institutions that want to advance the agendas of open access to scholarly information and the management and preservation of digital material. At HP we are excited by the role that DSpace can play as a vehicle for exploring and developing standards, and for ongoing research in digital asset management, archival, and preservation systems. Together we anticipate that DSpace will play an important role in the future of academic libraries and archives, and we look forward to productive collaboration with other institutions in this area.

Acknowledgements

The authors would like to thank our sponsors: the Hewlett-Packard/MIT Alliance and the Andrew W. Mellon Foundation. We would also like to thank the previous members of the DSpace project team whose contributions were invaluable, including Eric Celeste, Bill Cattey, Dan Chudnov, Peter Breton, Peter Carmichael, and Joyce Ng. Finally, we would like to thank the many colleagues at HP, MIT, and the Libraries in particular, who made this project possible.

Notes

[1] Berkeley Standard Distribution License, <http://www.opensource.org/licenses/bsd-license.php>.

[2] Particularly the work described in Arms, <http://www.dlib.org/dlib/July95/07arms.html>; Kahn and Wilensky, <http://www.cnri.reston.va.us/home/cstr/arch/k-w.html>; and the FEDORA project, <http://www.fedora.info>.

[3] METS information is available at <http://www.loc.gov/standards/METS>.

[4] SourceForge.net, <http://sourceforge.net/projects/dspace>.

[5] DSpace, <http://dspace.org>.

[6] Downloadable software developed by researchers at HP Labs, <http://www.hpl.hp.com/research/downloads/>.

[7] Since the system is written in java in can, in theory, run on other platforms than UNIX but this is untested by the DSpace development team.

[8] Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), <http://www.openarchives.org/OAI/openarchivesprotocol.htm>.

[9] OAICat can be found at <http://www.oclc.org/research/software/oai/cat.shtm>.

[10] See the Handle System® <http://www.handle.net>.

[11] See <http://web.mit.edu/oki> for more information about the Open Knowledge Initiative.

[12] See <http://www.ocw.mit.edu>, for more information about OpenCourseWare.

[13] See <http://web.mit.edu/about-mit.html> for MIT's mission statement.

[14] See the arXiv.org e-Print archive, <http://arxiv.org/> at Cornell University for information about the arXiv project.

[15] See <http://www.dspace.org/mit/plan.html> for the MIT Libraries' DSpace business plan.

Top | Contents
Search | Author Index | Title Index | Back Issues
Previous Article | Next Article
Home | E-mail the Editor

D-Lib Magazine Access Terms and Conditions

DOI: 10.1045/january2003-smith