DSpace: A Year in the Life of an Open Source Digital Repository System

MacKenzie Smith[1], Richard Rodgers[1], Julie Walker[1], and Robert Tansley[2]

[1] MIT Libraries, 77 Massachusetts Ave, Cambridge, MA 02139
{kenzie,rrodgers,jhwalker}@mit.edu
[2] Hewlett-Packard Laboratories, One Cambridge Center, Cambridge, MA 02142
robert.tansley@hp.com

Abstract. The DSpace™ digital repository system was released as open source software in November of 2002. In the year since then it has been adopted by a large number of research universities and other organizations world-wide that need a digital repository solution for a number of content types: research articles, gray literature, e-theses, cultural materials, scientific datasets, institutional records, educational materials, and more. The DSpace platform and its various applications are becoming better understood with experience and time. As one result of a recent meeting of the DSpace user community, we are now venturing into the territory of broad, community-based open source development and management, and gaining insights from the experience of the Apache Foundation, Global Grid Forum, and other successful open source projects about how to build open source software for the digital library domain.

Introduction

DSpace™ is a free, open source software platform for building repositories of digital assets, with a focus on simple access to these assets, as well as their long-term preservation (to help ensure access over very long time frames) [1]. It was originally designed with a particular service model in mind: that of institutional repositories of research material, and particularly research articles, which are produced by academic research institutions [2]. The idea was that institutions of all kinds could and should accept stewardship responsibility for their intellectual research output, for its widespread and long-term access. This is related to, but not synonymous with, the Open Access movement[1], since while many of the institutions using DSpace have free access to their assets as a goal, the platform itself does not assume that assets it stores will be made available for free.

DSpace was originally designed by developers at the MIT Libraries and HP Labs to be a breadth-first system with functionality to capture, describe, store, and preserve digital content, which adopters could download and install with minimal configuration and customization [3]. This decision was made for two reasons: to test the value of archivally-oriented digital asset management systems to the research university community without the need for extensive technical development, and to get a system

[1] E.g. the Budapest Open Access Initiative. http://www.soros.org/openaccess/initiatives.shtml

R. Heery and L. Lyon (Eds.): ECDL 2004, LNCS 3232, pp. 38–44, 2004.

out to the open source development community that was "good enough" to get things going and foster wider debate about the many technology choices involved.

Since its launch as an open source project in November of 2002, DSpace has undergone widespread adoption in several communities, and is starting to undergo active development by an open source developer community. This process of going from research to a public production release 1.0 on SourceForge, and then to a platform that is being developed by a large group of software developers representing both the original target audience and others who were not foreseen is an interesting story. It is our belief that the academic research community who often create open source projects for very good reasons don't necessarily understand the implications of the open source model or the long-term issues it raises. Our experience with DSpace is both atypical of many of the successful open source projects and also instructive to other research projects with a goal of becoming successful open source projects as their long-term business plan.

As a research project, it was the goal of the MIT Libraries or HP neither to productize DSpace, nor to continue to provide sole support and development of the platform going forward. Both organizations continue to work on the platform, in different areas and for different reasons, and we are committed to making sure that the platform has a viable and sustainable model for its ongoing development and adoption. That means ceding a large degree of control in order to gain the long-term vision of a self-sustained tool that we can all leverage to our best advantage.

This article attempts to provide enough context for DSpace to explain its origin and goals, to report on what has happened during the first year of its life as an open source project, and to attempt to divine the future of its transition to the next phase.

Background

The DSpace project was born out of a need voiced by faculty to the MIT Libraries to create a scalable digital archive that preserves and communicates the intellectual output of MIT's faculty and researchers. At the Institute, there is a growing body of digitally born materials representing significant intellectual assets that require stewardship. In addition to the more traditional text-based research output such as preprints and working papers, these assets include audio files, videos, datasets, software simulations and more. Faculty members often post their work on personal or departmental websites, but increasingly have become concerned about the sustainability of that solution. DSpace offers faculty and researchers a professionally managed archive that allows easy accessibility to their scholarly work.

Recognizing that the problems DSpace seeks to address are not unique to MIT, the MIT Libraries and HP Labs envisioned a federated repository based on a common set of institutional repository standards for interoperability. Interoperability would make available the collective intellectual resources of the world's leading research institutions. Further, we opted from the beginning to make the software entirely open source with the hope that a community of users and developers would emerge beyond the original MIT and HP team to contribute to the maintenance and enhancement of the code base over the long term.

The DSpace Federation

In January 2003, the MIT Libraries embarked on a project funded by the Andrew W. Mellon Foundation to work with seven other research universities to begin the process of building a collaborative federation of institutions running DSpace. Each of the seven universities installed DSpace and tested the adaptability of the system to their university environment. Our goal was to learn from these implementations and to share lessons learned with a wider DSpace Federation.

From the time the system was released as open source in November of 2002, uptake of the system has extended well beyond the original Federation project partners. DSpace has been adopted by a large number of research universities and other organizations that need a digital repository solution for a number of content types. These universities have evaluated DSpace's functionality and are further developing it to meet their needs. As the moment the software has been downloaded nearly 10,000 times; over 125 universities are investigating it for use in their university environment; and at least 20 universities are running production DSpace systems.

With interest in and use of DSpace mounting far more quickly than was originally anticipated, the set of institutions participating in the DSpace Federation project made the strategic decision to expand the final project meeting to include all institutions currently using DSpace and shift the purpose of the gathering to an open user group meeting, which was held on March 10-11, 2004. Approximately 120 people attended the sold-out meeting, representing 50 institutions, including universities, government agencies, and corporations, from 10 different countries. Members of the user community shared their DSpace experiences and plans, through which we learned that the DSpace platform is being put to a variety of uses: primarily to create institutional repositories of research publications and other material, but also for other applications (e-thesis repositories, learning object repositories, e-journal publishing, cultural material collections, electronic records management, and so on).

Within the UK, we already are beginning to see the diversity of purposes to which DSpace can be applied. The DSpace@Cambridge project, a joint collaboration between Cambridge University Library and MIT Libraries, aims to implement an institutional repository for scholarly research, but also is exploring the use of DSpace for administrative records and learning objects. Edinburgh University chose the DSpace platform for its Theses Alive! project, which aims to produce an OAI-compliant repository for the creation and management of e-theses and pilot it as a national service. Programmers at Edinburgh have developed an add-on module for DSpace that includes a supervised workspace for theses creation, supervision administrative tools, and a submission system for theses metadata collection. Glasgow's DAEDALUS project is piloting several open source institutional repository solutions and has opted to deliver a range of distinct open access services supported by complementary software platforms (one of which is DSpace) that optimally meet Glasgow University's needs for specific collection and digital content types. For the international community, it is also relevant to note the work done by the Université de Montréal's Érudit project to translate DSpace into French, work that has provided important lessons for customizing DSpace for local language. Other institutions in non-English speaking countries are now working to translate the system into local languages, and have identified general internationalization as an important goal for the future.

DSpace, and institutional repositories in general, are proving to be a high-value, long-term vision, but are still very much works in progress. Universities are setting their own policies to define what an institutional repository service means in the context of their university environment. Seemingly straightforward questions such as what types of file formats or content will be accepted and who is authorized to submit materials to DSpace quickly become complex when long-term implications for digital preservation and stewardship are considered.

Building collections of digital content, particularly scholarly research content, has proven to be another challenge universities consistently grapple with when implementing institutional repositories. Many of the DSpace projects around the world are grant funded or have limited resources and are under pressure to prove the value of the service, often measured (rather simplistically) through the number of items in the repository. DSpace was designed with a decentralized web submission interface that allows research communities to contribute their own items and metadata. This paradigm shift has been a novel and attractive aspect of the service but has meant that library staff has had to become proficient marketers, carefully positioning the service to meet user needs. Publicity and promotional activities help raise initial awareness among potential users but targeted communications with highly tailored marketing messages often are what persuade them to become submitters.

Open Platform – First Steps

The DSpace software released as version 1.0 into open source embodied use-cases derived from an analysis of needs within the MIT scholarly community viewed through the lens of the library. Yet this begged an important question: to what extent did these use-cases reflect the needs of institutional repositories generally? Rather than undertake a systematic survey or study, the expectation was that those who evaluated or adopted the software would provide an answer in the form of reworking the software itself to suit local purpose. The evolution of the DSpace platform would then consist of a rational assimilation of this work into the centrally managed code repository. Our biggest concern was the possibility of fragmentation or 'centrifugal' dissipation: that the platform would be pulled in too many directions, asked to do too many things, so that none could be done well. To prevent this, procedures were instituted to subject proposed contributions to a closely managed review process. Those of sufficient technical merit and deemed consistent with the vision of DSpace would be incorporated; the rest would reside as localizations of the platform outside its management.

The first year produced relatively few contributions, given the size and interest level of the adopter community. This was not due to a shortage of ideas, however: the mail lists and other forums were filled with use-cases and other expressions of need exceeding the 1.0 platform capability. Analysis of this situation revealed several factors at work: (1) The process of adopting DSpace could be lengthy and involved, and technical rework was often put behind such tasks as formulating a sustainable business model, developing service guidelines, or building awareness and buy-in from depositors. This had the effect of pushing software development considerations out of an early time frame. (2) Many of the potential adopter institutions lacked the technical resources required to undertake significant software development. (3) Architectural

limitations in the implementation of the platform made certain kinds of modification difficult to do. (4) Perhaps most interesting, however, was the perception that the platform, although distributed freely in source code form, was an immutable offering, much like commercial software product offering. There are many reasons why this perception took root, including the fact that its initial development cycle was 'closed', and that in order to build awareness of the platform it was 'branded' as an MIT/HP-sponsored effort, rather than an outgrowth of a community-driven process.

To address this perception, DSpace development was deliberately steered in the direction of the needs of the nascent community of users. The functional requirements of the next major release of the platform, 1.2 (1.1 basically represented the completion of the original research project agenda) were culled from postings to the DSpace lists, and from other discussions and surveys eliciting adopter feedback. In this way we hoped both to realize and to convey the community-centric nature of the DSpace platform. And to the degree that this additional functionality will remove barriers to adoption, the plan is proving successful. Yet since the bulk of the development effort was still concentrated within MIT/HP, it also is having the opposite effect – that of reinforcing the vendor/consumer dichotomy it was intended to overcome.

On the technical architecture front, the analysis of limitations has produced a roadmap for a new design direction, DSpace 2.0, which will address several key shortcomings of the current architecture: (1) Functional modularity coupled with the use of stable, well-defined APIs for their use will promote the development of independent implementations by DSpace adopters. This will substantially alter the concept of DSpace as a closed body of code, replacing it with the concept of a software framework, within which myriad implementations may coexist. (2) A refactoring of the presentation layer will enable much simpler alteration of UI without complications elsewhere in the code. (3) A much cleaner representation of content and associated metadata as a self-contained archival information package (AIP) will facilitate interoperability and maintenance of a DSpace repository.

From Code to Community

One important lesson we learned was this: to build an open source community, it is insufficient merely to publish a body of code as open source, even on commercially-friendly licensing terms (BSD[4]), and wait for a community to coalesce. Achieving true community requires the transformation of users who are initially consumers into stakeholders. We are examining several successful open source initiatives, such as the Apache Software Foundation, the Global Grid Forum, and the Eclipse Foundation, and, together with the user community, are formulating a plan for the DSpace platform. Among short-term objectives are: (1) expansion of the core set of developers to include those outside the initial circle of researchers. (2) Articulation of a clear process to encourage further enlargement of the developers' group. In most open source models, the existing group invites new developers, and functions as a project management board. (3) Recruitment of contributors to the platform on many other levels, from requirements definition to documentation, testing – indeed all aspects of platform maintenance and evolution. (4) Improved communication channels. Two goals are involved here: first, to produce greater *transparency* in the process of platform development we will need better ways to expose the deliberative steps involved. A developer-focused mailing list is one frequently adopted technique to achieve this.

Second, there need to be more flexible and accessible opportunities to become involved in development issues. Wikis and other semi-structured discussion tools can serve this purpose.

In the longer term, it will be important to establish or join forces with an independent not-for-profit entity (e.g. a 501(c)(3) corporation[5]) to be charged with stewardship of the software, and to possibly assume ownership of the intellectual property (copyright, trademark, license, etc.). Issues of financial support and governance models will be foremost in choosing a model – e.g. does financial contribution confer special privileges with respect to platform development? We hope to address these issues carefully, while proceeding quickly on the short-term agenda. Throughout this process, what is paramount to communicate to the greater body of adopters is that the continued evolution and in fact the very existence of the platform will depend upon a collective effort, not on the beneficence of the founding institutions.

Conclusion

As stated earlier, while it is not the current aim of the MIT Libraries or HP to build a commercial product with DSpace, neither was it our aim to prevent that from happening at all. We wanted to understand what it would take to build a useful digital archival repository: to test the technologies involved, to have a platform to explore service models like institutional repositories, and to have a platform for ongoing research in important areas such as digital preservation, Semantic Web techniques for metadata management, persistent identification schemes, and open access-friendly DRM systems.

In order to achieve our goals for the DSpace platform it is vital that it become a successful open source project with an active community of developers far beyond MIT or HP. That can only happen if the platform is useful to a critical mass of organizations that can provide the resources to do this work. We also expect that development of the platform will reveal a range of necessary standards – for interoperability, for rights managements, for identification of content and the people accessing it, for content discovery and preservation, for the metadata to support all of this, and more. The future DSpace Federation organizational home will provide the governance to make sure that everyone's goals for the platform are met, and hopefully to foster its adoption by a range of organizations in many sectors. The research community who we represent is but one potential adopter of this technology, and we believe that by leveraging the expertise and resources of other sectors, ours will ultimately benefit in ways that have proved elusive in the past.

The promise of open source for projects like DSpace to the digital library community are obvious, if it's successful. But there are some barriers to success. Many institutions lack the resources to deal with complex applications like DSpace on a technical level – they require support to install, configure and customize it for their local needs, and to maintain it over time. The open source world, with a few exceptions (most notably Red Hat for the LINUX operating system), doesn't provide models for such support and assumes that adopters have the necessary local expertise. The second barrier to success is in sustaining the developer community that will ensure the platform's continued usefulness over time. The research library community, who have been the primary adopters of the DSpace platform so far, do not, by themselves, have the resources themselves to sustain DSpace indefinitely. They have technical exper-

tise, to be sure, but it is typically over-stretched. They often cannot dedicate programmers to work on an open source platform without external support (usually for a grant-funded project of a year or two). It is possible that DSpace could survive on that basis, but risky. If the digital library community can share control of the platform with other sectors, particularly commercial and governmental sectors, then many more resources can be brought to bear to the problem.

So why are research libraries motivated to get involved in open source projects like DSpace? To learn more about how it really works, to better fulfill their mission, because commercial offerings are too expensive and often inadequate. Why are research libraries not getting involved? There is a noticeable tendency among managers to treat open source software as if it was commercially supplied. The owning organization is the "vendor" and the "products" can be comparatively evaluated and judged good or bad accordingly. The problem with this approach is that where open source software is concerned we are all, collectively, "the vendor". Or rather, there is no vendor to negotiate with, and if the product doesn't meet local needs then it can be made to do so. There is a corresponding tendency among library adopters of open source software to feel faint obligation back to its source – the software is just a product that happens to be free. But open source software certainly does cost its adopters something: the staff time to configure and maintain it without a formal support contract (typically), and the more nebulous moral obligation to provide some value in return for this free good. If open source software works at all then it's because those who benefit from it also contribute to it in some way: functionally, technically, or monetarily. Our community has much to learn about ways in which is can contribute to these efforts other than as grateful, but silent, adopters.

We have looked for inspiration to existing open source projects and organizations: obviously LINUX and the Apache Foundation, but also the Global Grid Forum and CNRI. Each of these organizations has some model for sustaining open source software but they're all different. Undoubtedly there are many others that we have not yet had time to identify and investigate. Which one is the most relevant to applications like DSpace? Which to the communities that created it? Which to the communities who are now adopting and improving it? Clearly there are many, many issues still to be addressed, and we hope that our experience in some way informs the understanding of the open source promise to and contract with the digital library community.

References

1. DSpace: An Open Source Dynamic Digital Repository. Smith, et al. D-Lib Magazine 9:1, January 2003. http://www.dlib.org/dlib/january03/smith/01smith.html
2. Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. Clifford Lynch. ARL Bimonthly Report 226, February 2003. http://www.arl.org/newsltr/226/ir.html
3. DSpace Internal Reference Specification. Bass, et al. March 2002. http://dspace.org/technology/functionality.pdf
4. BSD is the "Berkeley Software Distribution" license originally written in 1979 at the University of California, Berkeley for their open source unix software http://www.opensource.org/licenses/bsd-license.php. It is considered one of the "classic" open source software licenses, and the most commercial-friendly since it allows commercial development using the open source code.
5. 501(c)3 organizations in the U.S. are legally-recognized, registered non-profit organizations