2016

Choosing a Repository Platform: Open Source vs. Hosted Solutions

Hillary Corbett
Northeastern University, h.corbett@neu.edu

Jimmy Ghaphery
Virginia Commonwealth University, jghapher@vcu.edu

Lauren Work
Virginia Commonwealth University, lawork@vcu.edu

Sam Byrd
Virginia Commonwealth University, sbyrd2@vcu.edu

© The Author(s) 2015. This is the authors' version of a work that was accepted for publication in Making Institutional Repositories Work (Charleston Insights in Library, Archival, and Information Sciences), Burton B. Callicott, David Scherer, Andrew Wesolek (eds), Purdue University Press: 2016. The final publication is available at http://www.thepress.purdue.edu/titles/making-institutional-repositories-work

Choosing a Repository Platform: Open Source vs. Hosted Solutions

Authors: Hillary Corbett, Jimmy Ghaphery, Lauren Work, Sam Byrd

Introduction

Platform selection is a concept that will be familiar to many who work in libraries, regardless of whether they have worked with an institutional repository. Selection and implementation of a new integrated library system (ILS) or discovery platform are experiences that most library staff will generally encounter more than once in their careers, and they are processes that typically represent a significant, long-term time commitment for staff across the organization. The stakes are high because so many library employees' day-to-day work involves active and extensive use of the system that is chosen. Because of this common experience, it naturally follows that library staff tasked with choosing an institutional repository platform may approach the job with trepidation. But in reality, the selection process doesn't have to be as time-consuming or fraught with anxiety. (Indeed, a common pitfall may be to over-plan for the process.)

While it's essential to include representatives of different areas of expertise, the group tasked with selection can be fairly compact. This will help the process move more smoothly. Who should be included in this group? If there is an existing repository, its manager should be involved, of course. Staff from metadata and systems units should also be included. Even with a hosted platform, where no on-site technical expertise would be needed, the systems representative will likely be best able to evaluate its architecture and interoperability. Someone with an archival background can also provide valuable perspective on the preservation aspects of the repository platforms under consideration. Your web developer or user experience expert can be very helpful in evaluating interfaces and their potential customizations. Above all, the repository must be useable. It can have great metadata support and elegant architecture, but if the interface is clunky, no one will use it. A team member who knows how users interact with the library's other online resources is essential. Finally, you may also wish to seek input from a power user of your current repository, or someone who is likely to be an active user of a repository under development. If including them during the selection process isn't feasible, such users should certainly be asked to help later with usability testing.

Your library may already have an existing repository, but try to evaluate prospective new platforms independently of whether or not they are "better" or "worse" than your current platform. In many ways, a new platform will likely just be different - and that's going to be a combination of positive and negative. Of course, it's important to consider your current platform in the context of how you will migrate its contents! But you've already made the decision to move to a new platform - strive to evaluate your choices on their own merits. The goal in your selection process is to compare new platform with new platform, not new platform with current platform (or with the absence of a platform, if you don't currently have a repository). If your library already hosts a repository and you're looking for a new platform, you should certainly make a list of your current platform's pros and cons - but don't let them influence your process too much or get bogged down with too much discussion of the current platform. Likewise, keep in mind that platforms are constantly under development, and specific features you note as absent or less well-developed may well be slated for future releases. Most importantly, remember this evaluation is not a mere side-by-side comparison, but needs to be tied to your institution's repository goals and ambitions.

While this chapter discusses selection of a locally hosted, open-source system (DSpace/Fedora) versus a cloud-hosted, proprietary system (Digital Commons), it is important to note that these examples are merely illustrative. Libraries have a range of choices for repository software that includes open source and proprietary in any number of support environments, and exemplary repositories are flourishing on a variety of systems, both open source and proprietary. This chapter focuses on the differences between proprietary and open-source solutions, but also demonstrates how and why libraries choose a repository system. In writing about this process, we realized that it was important to acknowledge that there are two different audiences for this chapter: those who may just be starting out with building a repository at their institution, and those with an established repository who are considering a platform change. Thus, this chapter addresses the challenges and opportunities of platform selection in both circumstances.

Selective Literature Review

The library literature regarding open source software has dealt with a variety of systems, including integrated library systems (ILS) and repository platforms. Pruett and Choi’s (2013) article comparing select open source and proprietary ILS software includes a thorough review of previous research, including welcome background from fields other than library science. Palmer and Choi’s (2014) descriptive literature study is also an important touchstone for an understanding of previous research on library open source software. In this review, the authors found that almost 35 percent of the library literature regarding open source has dealt with digital repository software, and posit that this concentration is largely due to a preponderance of open source repository platforms (DSpace, Fedora, EPrints). Indeed, the repository market is almost an opposing image of the open source ILS market since open source solutions have defined repository solutions from the outset.

Library literature concerning the choice between open source or proprietary repository platforms reflects the multi-faceted and unique circumstances that individual institutions face. Burns, Lana and Budd (2013) reflect this reality in the conclusion of their survey of institutional repositories, stating that “the most important lesson learned from this survey is that not all institutional repositories are alike” (Discussion, section 5, para 1). Though widely applicable evaluation methodologies and parameters for choosing an institutional repository are well documented (Rieger 2007; Fay 2010; Giesecke 2011), final decisions for open source versus proprietary platforms are most often unique to the circumstances of each institution and emerge from university-level needs assessments. Common factors cited in the case studies for choosing proprietary solutions include costs of technical infrastructure and staffing; the need for swift implementation to allow for a focus on repository population and promotion; interface branding and customization; electronic publishing options; and online discoverability of scholarly research

(Mandl & Organ 2007; Bluh 2009; Younglove 2013). Libraries that select open source repository platforms also note customization as a positive factor, but include extensibility, flexibility to ingest varied formats, and interoperability (Marill & Luczak 2009; Fay 2010). In line with these cost-benefit issues of open source, Samuels and Griffy's (2012) case study in evaluating open source publishing solutions includes a comparative methodology that includes total cost of ownership.

Salo's tongue-in-cheek essay "How to Scuttle a Scholarly Communications Initiative" (2013) is required reading, both for its insightful look into library culture and its very well-developed bibliography for anyone interested in starting or improving a scholarly communication program. In discussing platform choice, Salo encourages usability and beta testing as well as reaching out to colleagues who are current or former users of the systems under consideration. Salo makes her point about the pitfalls of focusing solely on platform without consideration of the larger scholarly communication goals of the organization in a particularly humorous manner: "It is particularly important to fixate on a software package before the initiative's mission, milestones, and workflows have been decided….to maximize the discrepancies between necessary work and the software's capabilities" (p. 3).

Virginia Commonwealth University: From Open Source to Proprietary

Virginia Commonwealth University (VCU) launched a DSpace instance in 2007 as a platform to support its electronic theses and dissertations (ETD) program. All systems and database administration, server maintenance, and application support were handled by library technical staff. There were no additional staff allocated for the ongoing support of the repository. The initial installation and support were carried out by the Web Systems Librarian, who relied heavily on the DSpace-tech listserv[1] for support and advice. Shortly after launching DSpace, the library sought clarification of its goals for the repository. A Statement of Direction was developed that intentionally limited use of DSpace to deposit of ETDs, for several reasons: anticipated difficulty in supporting an expanded DSpace repository, environmental scans of difficulties that other fledgling repositories were facing, and a sense that focusing on digitization of local library collections would yield greater impact.

Once DSpace was installed and launched, support did not entail any significant work beyond routine operating system patches. The ETD collection grew without incident. In 2010, VCU's Web Systems Librarian, who served as the lead support person for DSpace, left the university for another position. It was not possible to find a replacement who had the same level of DSpace expertise, which was problematic due to an anticipated need to upgrade both hardware and software. Migration of embargoed ETDs while preserving their security was of particular concern. While VCU had previously received help for some issues on the DSpace-tech listserv, this type of assistance was not always consistent or sufficient to support what was becoming a larger and more mission-critical collection of ETDs. For all of these reasons, the library contracted with a vendor to provide support services specifically for upgrading the software.

[1] https://lists.sourceforge.net/lists/listinfo/dspace-tech

This upgrade process was a significant task. It included vendor support in testing the new version on a hosted sandbox server as well as local work in writing custom SQL code to move retrospective embargo data to new database fields. After the successful migration, the decision was made to continue vendor support. On January 9, 2014, it was announced on the DSpace-tech listserv that, consistent with the DSpace Software Support Policy[2], the version of DSpace being used at VCU would no longer be supported with security patches. Even though VCU had already made the decision to move to Digital Commons at that point, issues with local upgrades of DSpace were one of the factors that encouraged us to move to a cloud hosted solution. While VCU did face some technical challenges with DSpace, we were by no means dissatisfied. An official software support policy is an excellent step toward keeping software moving forward, and the software was very stable with only minor issues. We achieved this consistency of performance without major staff investments. And like other enterprise-level library software, DSpace was not unique in requiring significant effort in testing and deploying upgrades.

Meanwhile, the library had been making modest steps toward expanding the scope of the repository. In 2013, two collections were published on the DSpace platform: *British Virginia*, a peer-reviewed series of scholarly editions from and about the Virginia colonies, and an annual series of undergraduate research posters. Both of these projects engaged external departments at VCU who saw great benefit in partnering with the library in these publishing endeavors. The field of scholarly communication and library publishing had likewise shifted dramatically since our cautious 2007 assessments, with a number of successful models.

The desire to expand the library role in publishing was also surfacing as a new need. Based on our own research, and previous experiences running DSpace, we felt that DSpace would not be adequate as a journal publishing platform. As such, if we remained on DSpace for our anticipated repository growth, we were also looking at implementing another system to support journal publishing such as Open Journal Systems (OJS). We considered various combinations of local and hosted implementations of DSpace and OJS. We did find the open-source virtues of these systems, and the natural alliance of open source and open access, to be compelling. However, after much discussion across the organization, and against the backdrop of recent successes with migrating our other major library systems to the cloud, we decided that Digital Commons was our best path forward to quickly meet our ambitions.

Beyond the vendor-supported cloud platform and its integrated repository and publishing systems, there were a number of other enticing features of Digital Commons that led to our decision to migrate. We were drawn to the marketing and outreach features of Digital Commons and were excited about features such as automated author notifications, federated networking of all customer content, and search engine optimization. These functions seemed difficult to reproduce with open-source solutions, especially given VCU's systems staffing. And because of recent experiences with other cloud-based systems, we knew that the process of implementing

[2] https://wiki.duraspace.org/display/DSPACE/DSpace+Software+Support+Policy

new releases would likely come with less overhead than we were used to on a locally supported system.

VCU's implementation of Digital Commons was rapid, enabled by a number of factors. During a two-month period, design decisions and IR policy outlines were established – library administration wanted to move the project along quickly, and a task force was established that helped accelerate this progress. VCU signed its contract with Bepress at the beginning of February 2014, the repository went live in March, and accepted its first submission in the new system in April.

The migration of VCU's data from DSpace highlighted structural differences between the two systems and the importance of portability of repository data. In certain areas we ran into difficulty reconciling differences between the two platforms. One had to do with how supplemental files are handled; another was representation of special characters and diacritics in the metadata. The actual handling of the Dublin Core metadata was significantly different for each system, particularly for the date and creator fields. A number of bulk loads, revisions, and finally some targeted manual editing were needed to complete the project. Bepress customer support was extremely helpful during this process, but in the final analysis it was our responsibility to migrate, test, and accept data.

There are a number of features from DSpace that we certainly miss. We obviously do not have direct database access and must depend upon the vendor for certain reports, including quarterly backups. Many things require vendor intervention, such as setting up a new collection. Fortunately, Bepress provides an exemplary level of customer support to turn our requests around quickly. We have embraced the limitations of the user interface design templates with an understanding that common design patterns across all customer sites enhance the ability for agile product improvements.

We have been impressed thus far with new features and strategic directions of Bepress, including more intentional support for datasets and images. A few other qualities of Digital Commons have also been affirming our platform decision. We have seen initial evidence that the author notification and search engine optimization features that appealed to us in the selection process also appeal to our users at VCU and are fostering greater acceptance of the repository. The road toward establishing mature repository and publishing services, however, is long, and we are admittedly at the start of the journey. Our current confidence in and excitement with the Digital Commons platform is enabling us to offer these services to the university community in a way that seemed out of reach to us before.

Northeastern University: From Proprietary to Open Source

As an early developer of an institutional repository, the Northeastern University Libraries have perhaps had a wider range of experience with IR platforms than many institutions. Northeastern began building its first repository instance in 2004, in a development partnership with Innovative Interfaces. The repository, called IRis, was launched in 2006 using Innovative's Symposia platform. While a proprietary system, Symposia was mounted locally and required a significant commitment from library staff. In 2009, the library decided to move to a hosted repository platform in order to free up staff to work on other strategic priorities, and migrated to Bepress's Digital Commons solution.

A hosted solution is an excellent long-term option for many institutions that do not have the local resources to develop and sustain a repository built using open-source software. A hosted solution can also serve as a first step during the time that a local repository is being developed. However, the amount of time needed to develop the local platform may end up being significantly greater than originally anticipated. We found this to be true at Northeastern. When Northeastern transitioned to Digital Commons at the end of 2009, we already expected that it would be a medium-term solution until the library had the resources to build and support a Fedora-based repository. At the time of this writing, in Fall 2014, our Fedora-based Digital Repository Service (DRS) has at last entered a soft-launch phase after two full-time staff years of concerted effort from our web developers. Full release of the DRS is slated for January 2015.

Northeastern chose to model the DRS after Pennsylvania State University's Fedora- and Hydra-based ScholarSphere repository.[3] Converting the ScholarSphere engine for our purposes and removing its existing dependencies was challenging, although the developers at Penn State extracted functionality from ScholarSphere into a new open-source web application called Sufia[4], which our developers were able to make use of. Another challenge to development of the DRS was the need to support a prototype model that had gone into production earlier than planned in order to support immediate on-campus needs that could not have been met by the Digital Commons–based repository.

Our goal when developing the DRS was to have all our digital assets—faculty-authored materials, electronic theses and dissertations, learning objects, digital special collections, and archival materials—managed by a single architecture. Most importantly, a local repository, built with open-source software, gives an institution total control over its content and how it is organized and displayed. Open-source software like Fedora offers flexibility for local customization to an extent not possible with a hosted platform with hundreds or thousands of clients. With a locally developed repository, it becomes easier to meet the specific needs of local users, as opposed to offering a product that has been developed to meet the more commonly encountered needs of the average repository user. The types of materials being deposited in the repository may also drive development – at Northeastern, a department wanted to deposit large quantities of images directly from digital cameras, and have thumbnails automatically generated while preserving the original large files. We were able to customize the deposit interface to make this possible for them, and for future users with a similar need. Understandably, the providers of a hosted IR solution would not be likely to take on this type of customization work for a single client.

[3] https://scholarsphere.psu.edu/
[4] https://github.com/projecthydra/sufia

At institutions where the majority of IR deposits are PDFs, an “out-of-the-box” solution that requires little customization works very well. While its infrastructure can certainly accommodate other types of materials, the manner in which non-PDF materials are arranged and presented can be limiting. However, with an open-source solution like Fedora, another open-source tool like WordPress or Omeka may be used to create a “discovery layer” that exposes content from the repository in a manner that is more meaningful and appropriate especially for non-textual materials. We recently worked on such a project for a group on campus who wanted to store videos in the repository, but make them available through a site that could also present other content in a flexible interface. A WordPress instance was a good solution for this need, and created a strong use case for future projects. The ability to make use of a robust repository infrastructure while exposing content in non-“repository-like” ways will certainly serve to make the repository a more attractive solution for potential campus clients.

While choosing to build a repository based on open-source software offers many opportunities for development and customization, it also comes with challenges. Aside from the time and technology costs required to get the repository from day one of development to a full production instance, there are also important ongoing workflow considerations. With a hosted repository platform, the library pays for customer support as part of the annual maintenance fee. With open source, there are online communities of developers using the same platform who can offer advice, but bug squashing may definitely be more challenging.

Academic libraries sometimes have trouble retaining skilled developers, simply because they aren’t able to compete with the salaries offered in the corporate or startup worlds. The library should thus not assume that the person on staff who originally built their repository is going to be around to sustain active development. We found this to be the case at Northeastern; in fact, a significant amount of the repository development has been done by a student who has worked with us for several years. Repository developers should fully document their work as they go so that new staff can take over without interruption. Beyond the developer, the library should also have someone on staff to serve as the repository manager. While this role is necessary in any library with a repository, regardless of the platform chosen, in a locally hosted repository it is vital that the repository manager is able to be highly responsive, as there is no customer service staff elsewhere. At Northeastern we have moved from having the hosted repository managed by the scholarly communication librarian, who has other duties, to having a dedicated Digital Repository Manager for the DRS.

Ongoing support – both maintenance and continuing development – must not be overlooked as a cost when deciding to build a repository based on open-source software. The library must be able to *fully* support the repository – “adequate” support for such a significant and high-investment resource is not enough. Northeastern estimates that support for the DRS will equal 1.5 FTE - a full-time repository manager, and half of our senior web developer’s time. This is in sharp contrast to the staff necessary to support the Digital Commons–based repository: 0.25 FTE of the scholarly communication librarian’s position and a minimal amount of time (fewer than five hours per week total, on average) from two metadata staff.

For those who have worked with the repository at Northeastern, the transition from the Digital Commons platform to the open-source DRS is bittersweet. We are excited about the new opportunities for providing an increased level of customization for our users, and feel positive that the direction our repository's development takes will be entirely under our control. However, Bepress has been an excellent company to work with, and they made our use of Digital Commons a productive and important stage in the lifespan of our repository.

Conclusion

The VCU and Northeastern case studies are similar in their emphasis on choosing and implementing a repository platform to best serve local needs. Neither VCU nor Northeastern has found critical flaws in the systems from which they are migrating, and indeed both institutions' recent migrations were driven primarily by local priorities: VCU chose Digital Commons in response to an identified need to quickly provide enhanced repository and publishing services, and Northeastern decided to go open-source in order to offer greater customization and maintain control over content. These decisions echo the literature on repository platform selection: a locally supported open-source system allows maximum flexibility, whereas a proprietary system offers turnkey entry and support.

Both institutions' experience with migrating content from one repository system to another indicate an area for future research, as metadata and file standards can be implemented in different ways between systems. Planning for possible future migration is wise when considering how you implement and customize your current system. If repositories grow to include vast amounts of material, as we hope they will, it is not clear how existing migration strategies will scale.

It is also important to note that the distinction between open source and proprietary solutions has started to blur. Following the model in other industries, a number of commercial support services are available for open source systems, ranging from hourly vendor support to full software-as-a-service offerings. Likewise, some commercial firms provide a range of choices to libraries to either install software locally or host it offsite. In general, we feel that the repository system landscape will be brighter into the future as a result of competition between various service models. Finally, it cannot be overstated that the platform itself is not a panacea, but merely one component of the institution's repository service.

Works Cited

Bluh, P. (2009, July). *TCO and ROI: Assessing and evaluating an institutional repository.* Paper presented at the American Association of Law Libraries meeting, Washington, DC. Retrieved from: http://digitalcommons.law.umaryland.edu/fac_pubs/796/

Burns, C. S., Lana, A., & Budd, J. M. (2013). Institutional repositories: Exploration of costs and value. *D-Lib Magazine, 19*(1-2). doi:10.1045/january2013-burns

Fay E. (2010). Repository software comparison: Building digital library infrastructure at LSE. *Ariadne, Issue 64*. http://www.ariadne.ac.uk/issue64/fay

Giesecke, J. (2011). Institutional repositories: Keys to success. *Journal of Library Administration, 51*(5-6), 529-542. doi:10.1080/01930826.2011.589340

Mandl, H. E. & Organ, M. K. (2007). Outsourcing open access: Digital Commons at the University of Wollongong, Australia. *OCLC Systems & Services - International Digital Library Perspectives, 23*(4), 353-362.

Marill, J. L., & Luczak, E. C. (2009). Evaluation of digital repository software at the National Library of Medicine. *D-Lib Magazine, 15*(5-6). doi:10.1045/may2009-marill

Palmer, A. & Choi, N. (2014). The current state of library open source software research: A descriptive literature review and classification. *Library Hi Tech, 32*(1), 11-27. doi:10.1108/LHT-05-2013-0056

Pruett, J. & Choi, N. (2013). A comparison between select open source and proprietary integrated library systems. *Library Hi Tech, 31*(3), 435-454. doi:10.1108/LHT-01-2013-0003

Rieger, O. Y. (2007). Select for success: Key principles in assessing repository models. *D-Lib Magazine, 13*(7-8). doi:10.1045/july2007-rieger

Samuels, R. G. & Griffy, H. (2012). Evaluating open source software for use in library initiatives: A case study involving electronic publishing. *portal: Libraries and the Academy, 12*(1), 41-62. doi:10.1353/pla.2012.0007

Salo, D. (2013). How to scuttle a scholarly communication initiative. *Journal of Librarianship and Scholarly Communication 1*(4), eP1075. doi:10.7710/2162-3309.1075

Younglove, A. (2013). Rethinking the digital media library for RIT's The Wallace Center. *D-Lib Magazine, 19*(7-8). doi:10.1045/july2013-younglove