18 On Implementing an Open Source Institutional Repository

James Tyler Mobley

In 2005, in an attempt to streamline the graduate thesis submission and publication process, the Graduate School at the College of Charleston in Charleston, South Carolina, entered a contract with ProQuest/UMI Dissertation Publishing to use the ProQuest ETD Administrator platform for students to submit their works and have them made available online. Prior to this agreement, paper copies were submitted and processed directly by the Graduate School, and copies were later sent to the College of Charleston Libraries for cataloging. With the removal of the paper component of these thesis submissions, the library suddenly faced the question of how to pivot to preserving electronic copies and how to make them available for students and faculty in the long term. At the time, the library did not have a platform dedicated to electronic content created by the college's students and faculty. In fact, the library had almost no infrastructure to handle local storage of electronic content whatsoever.

The single "repository" of content within the library at this time was the Lowcountry Digital Library (LCDL). LCDL consisted of a CONTENTdm-based digital library created for the express purpose of digitizing and presenting cultural heritage materials from the Lowcountry region of South Carolina. This installation was hosted on servers maintained by the library acquired through grant funding. While this repository was not intended to house non-historical works, it was the only portal through which the library could effectively manage and present electronic materials, especially materials like theses that came with various access restrictions and embargoes. As such, a limited number of electronic theses received from ProQuest were processed by a library cataloger and placed into the Lowcountry Digital Library. Over the next few years, theses were sporadically added to LCDL, though a formal workflow was not in place.

In the spring of 2010, the Lowcountry Digital Library project initiated a migration from CONTENTdm to an open source digital library platform based on Fedora Commons. It was at this time that the library concurrently began considering options for an institutional repository (IR) system for the preservation and presentation of contemporary College of Charleston output like theses and other works by students and faculty. An institutional repository could potentially provide a long-term home not only for electronic theses but also the output of the college as a whole. Obstacles and considerations encountered during this search included a lack of dedicated funds, limited staff time and expertise, and uncertainty about the perceived demand for such a system.

Limited budget allocation for new software projects proved to be the greatest single obstacle while investigating options for an IR. The library has a limited annual budget, most of which goes to the collection and other essential expenses. There is not a dedicated fund for pilot software projects or other exploratory initiatives. Additionally, though it was agreed that a solution for handling theses and other content was greatly needed, library staff remained unsure about the potential use of and enthusiasm for such a system by the rest of the institution. We were hesitant to secure large amounts of money for an unproven concept that our students and faculty might not even want to use. Because of this, we knew from the outset that we would prefer an open source option if one existed with adequate features and community support.

In terms of staff capability, the library had four dedicated technology employees who could be considered relevant to installing and managing an IR. There were two Digital Services librarians, one Digital Scholarship librarian, and one server administrator, all of whom were tied up in a variety of tasks throughout the day supporting library technology as a whole. The library did not have its own internal IT, and campus IT are typically busy handling campus-wide applications and maintaining network security and coverage. Therefore, we needed a mostly packaged solution to implement. We were prepared to maintain existing systems, but we did not have the staff to dedicate to building new systems from scratch, especially alongside the digital library rebuild that was already in progress.

When we began to explore our options, one immediate thought was to leave the theses in the CONTENTdm installation that LCDL was leaving and continue the manual cataloging process. We could then add new materials from around campus into new collections in CONTENTdm. We would basically reset CONTENTdm as an IR. The library had, after all, already paid for a portion of the system, and it still functioned well overall. While this wasn't a popular idea, it was potentially at least more economical than others. However, after further evaluation, the ongoing annual maintenance fees for CONTENTdm and the looming cost of hardware replacements made even this option a substantial investment. Some investment would be necessary with whatever option we chose, but we preferred it at least be toward new and improved systems and services rather than maintenance on a process that already didn't work very well. With that in mind, we shelved this option and took a look at the upcoming digital library platform.

The new digital library platform is built on the Fedora Commons Repository, which offers a great deal of flexibility in storing and handling digital content. We briefly imagined placing new campus materials in this repository alongside LCDL's cultural heritage materials and accessing each set of content separately through different interfaces. This would allow us to keep heritage materials separate from general college materials within a single repository.

Unfortunately, staff expertise was limited when it came to separating pools of content within one Fedora Commons repository, and Fedora Commons does not include robust front-end features for access control or display. At the time, just the construction of the Lowcountry Digital Library as a Fedora Commons repository was proving difficult enough without adding another factor to the challenge. The digital library repository was thus abandoned as an option in favor of a new turnkey solution. Today, the Fedora Commons repository only houses cultural heritage materials for the Lowcountry Digital Library, and we do not have plans to further expand its scope in the near future.

Digital Commons from bepress quickly became a major option for us as a turnkey IR system once we abandoned hopes of leveraging existing internal systems. Clemson University had already purchased it for use with their campus materials, so it already had some buy-in among our state peers. Additionally, bepress provides a great deal of support to clients using Digital Commons. Dedicated support would be ideal for an institution like ours with limited staff.

Beyond support, Digital Commons offers a number of features that other solutions don't have by default and which would be very time-consuming to create in-house, including dedicated pages for faculty profiles and various custom theming options. User-friendly features like these made Digital Commons a very enticing option. It is very much a one-stop solution for an institutional repository.

Unsurprisingly, such a robust feature list and support system came with a cost. Given our previous concerns that the library and the college as a whole might not ultimately care for an IR in the long term, we could not commit to purchasing something like Digital Commons. Had we already noted an expressed demand for an IR system, our outcome might have been different. With this aversion to license agreements, we turned our gaze more firmly to the open source community.

While exploring digital library system options for the Lowcountry Digital Library migration, we had previously come across DSpace. DSpace is an open source IR application initially developed by MIT that, like Digital Commons, is meant to be mostly turnkey. DuraSpace, the same group that maintains the Fedora Commons repository, now curates it. For the purposes of LCDL and its largely visual cultural heritage materials, DSpace was not a perfect fit. However, when reevaluated in the context of institutional repositories, which is DSpace's intended use case, it immediately became a primary contender.

The open source DSpace immediately checked a large box in the cost department, at least in terms of licensing and contract fees. It would, of course, incur further costs in acquisition of server hardware and staff-hours, but it lacked a lump sum cost of entry. We could test, modify, break, and even soft launch DSpace on existing hardware with no consequence other than possibly wasted time.

However, open source alternately meant a higher barrier to entry in the form of technical expertise. DSpace is a Java-based application that, while very well documented and maintained, requires at least some personnel that can run, configure, and maintain such applications and the servers they need to operate. Additionally, all customization would have to be done in-house by existing staff.

Furthermore, as DSpace is not a vendor-hosted solution, storage and backup capabilities would have to be considered concurrently. We could not approach the campus with a solution that did not on some level promise long-term storage and preservation of its collective output. This would have been a consideration with any locally hosted option, however, so this was not a consideration unique to DSpace so much as to locally hosted solutions in general.

It became fairly clear when outlining an open source product alongside a proprietary system that the debate of cost was powerful but also misleading. We were not and are not able to lay down large sums of money for the purchase of new software for untested needs. However, the long-term cost in staff time and server hardware for an open source solution was not negligible either. Both solutions would incur costs, some more direct than others. In this case, the library already owned at least some existing hardware running various Web sites and services. Hardware acquisition and management would make even considering DSpace a difficult task for some institutions, but it fit well into our existing infrastructure. The prospect of hardware cost and maintenance was thus more palatable than software costs.

Despite these technical complications and storage needs, DSpace promised a huge list of features that rivaled a system like Digital Commons. User groups, access controls, batch item loading, search and discovery, and other features were available out of the box with some amount of configuration.

In addition to a wealth of native features, DSpace also had the benefit of a very active development community. In any investigation of open source software, one must consider the activity of the community surrounding it. As open source software does not comes with a license agreement for ongoing updates and support, it is vital to ascertain whether the application in question will see support from its own volunteer community over time. After all, you don't want your staff stuck maintaining abandoned code for years to come. Ultimately, we came to the decision that DSpace struck a healthy balance between cost and features for our initial trials.

After the decision to run DSpace as a pilot project for the IR, the application was briefly installed on a test server and run for staff demonstration and testing. After that period, the library was able to acquire new server hardware to provide adequate processing power and storage to this and other library projects. As previously mentioned, the library already maintained a number of servers hosting smaller, basic Web sites and some essential proprietary applications like interlibrary loan software. At this time, the library had gone a number of years without new hardware, and existing servers were both limited in space and nearing their end of life. The acquisition of new hardware allowed us to set up DSpace in a proper production environment. This was not an expected turn of events, but it greatly eased the process of implementation. This acquisition also benefited the aforementioned digital library project.

The actual installation process of DSpace on a virtual machine running Fedora Linux was fairly smooth thanks to documentation provided by DuraSpace. While the system takes a few extra steps to implement due to the nature of deploying Java applications to Web servers, the documentation provided a more than adequate guide for a user with intermediate server and application experience.

After this installation came a moderate amount of customization. How DSpace looks and operates is largely "up to you." There are a number of ways to approach your system, including two entirely different Web interfaces from which to choose. One is rendered in traditional JavaServer pages (the JSP interface), while Apache Cocoon powers the newer XMLUI interface. We opted for the XMLUI interface as it promised more flexibility and features moving forward. XMLUI, for example, was the first interface to have an integrated discovery interface built on Apache Solr.

DSpace also offered more than a few options for user authentication. The College of Charleston campus uses LDAP as a user authentication method, and DSpace provided an authentication plug-in to support LDAP by default. LDAP in conjunction with IP authentication fit very comfortably into our campus environment.

After the site was visually customized and allowed campus users to access it properly, we had to approach the issue of content organization. DSpace breaks content down into Communities and Collections. In our case, we decided to break college departments into Communities that could have their own Collections. Each of these Communities and Collections needed individually assigned access restrictions depending on the type of content.

Once this organization was complete, there came the matter of getting electronic thesis content, our initial test material, into the system. As a proof of concept, we batch loaded 32 electronic theses from ProQuest into the collection via DSpace's command-line batch processing interface. This content fit well within the native structure of DSpace's Community and Collection hierarchy, so we decided to move forward with a more streamlined submission and deposit process. Conveniently, DSpace supports the SWORD protocol for document deposits. ProQuest has recently implemented this protocol as well, so, after some communication with ProQuest technical support, electronic theses are now automatically deposited as complete items into DSpace by ProQuest. This workflow eliminates the process of retrieving a PDF and metadata file from ProQuest and manually processing it. Instead, catalogers can now simply check on the IR system when they receive notifications that new items have been deposited into the Electronic Theses & Dissertations Collection.

DSpace handily solved our initial use case for an institutional repository by giving our electronic theses a permanent home managed on our local servers. Now that we have an institutional repository in place, we will have to consider staffing allocation to handle the management of the application as well as workflows for new content from other sources. These details are currently under consideration by the library, and a faculty committee is drafting a formal policy for IR content. Additionally, the library will be hiring a dedicated metadata librarian, a large focus of whose role will be to directly oversee the institutional repository.

While these formal considerations and new positions are worked out, we have embarked on a few test projects using the system. We have worked with the College of Charleston Honors College on two projects that have allowed students to submit items via a submission form. These forms automatically submit items to the IR to appropriate collections. Both of these projects make use of the SWORD protocol alongside DSpace.

Small test projects like these have contributed to awareness on campus in small doses; however, the question of overall institutional interest in an IR remains. We believe that our IR built on DSpace can provide a home for the digital output of students and faculty at the College of Charleston.

However, we have to pursue faculty engagement to prove its value as a tool for preservation and presentation. Once we have formalized our internal processes, we will move forward with broader campus outreach.

What began as a question of where to house some PDF copies of electronic theses developed over the past few years into the construction of a potential home for the College of Charleston's scholarly output. The choice to go open source for this project let us experiment with new directions in our library systems without risking valuable annual library budgets or sinking too much staff time into developing homegrown applications. However, selection and implementation were only the beginning of a longer dialogue over the role of an academic library in preserving the collected academic output of its institution. At the College of Charleston, that dialogue is ongoing.