From Campus to Cloud

Planning and Implementing a Digital Repository Migration

By David J. Williams

My introduction to librarianship began with a varied career in technology, including work as a web developer and systems administrator. My academic interests encompassed electronic records and digital archiving. While completing my graduate studies in library science, I enjoyed the good fortune of supporting online resources in an academic library. These experiences proved invaluable, as after graduation I was offered the opportunity to serve as Digital Initiatives and Special Collections Librarian with William Paterson University's David and Lorraine Cheng Library, a position combining both technology project management and traditional archives practices.

As at many public universities, budgetary constraints limited the technical support available to academic departments, and campus information technology (IT) resources were dedicated primarily toward telecommunications infrastructure and network security. Prior to my arrival, the Cheng Library was tasked with developing digital collections, and initial projects were built using the OCLC CONTENTdm platform, a hosted digital asset management system offering customizable online displays. Over time the cost of this service became prohibitive, and the decision was made to develop a locally hosted repository capable of supporting institutional records and academic scholarship.

After considering several alternatives, the library adopted DSpace, a digital repository application initially developed at the Massachusetts Institute of Technology. Built according to archival principles, DSpace provides long-term preservation and access to digital content. Like many contemporary distributed applications, DSpace makes extensive use of middleware, primarily in the form of the Java SE platform, software that connects different applications, services, and systems within a single host or across multiple networks. This design allows for great flexibility but also introduces unavoidable complexity through a wide range of available options for database storage, operating environments, web servers, and search indexing. Complex technology projects require careful planning and testing, but time constraints, personnel changes, and limited IT support presented a series of challenges to the stability of this new service.

My first task as Digital Initiatives Librarian was to stabilize and secure the repository. This required a thorough assessment of the technical and design decisions implemented during the initial launch. The results revealed that existing collections were stable and accessible, but expanding to support future initiatives would be problematic. Conducting routine maintenance and applying security updates would also prove challenging due to limitations in the underlying hardware environment. Institutional repositories, digital or otherwise, must embody the archival principle of authority, and the structures in place for preserving collection materials needed careful reconsideration to address future reliability.

The next step involved gathering requirements and analyzing technical specifications. An additional benefit of distributed systems is portability, the ability to migrate components transparently to different locations and different supporting technologies. Adopting a flexible storage architecture, for example, with files managed separately from the host operating environment, facilitates portability by allocating additional space without introducing any changes to the application. Using this approach, digital asset storage could grow from 100 GB to 1000 GB without interruption. But

David J. Williams, MA, MLS is the Digital Initiatives and Special Collections Librarian at William Paterson University of New Jersey.

the issue of "insourcing" (rebuilding the service locally) versus outsourcing remained, as several certified digital repository providers offering a wide range of technical solutions were available. After extensive research, the results were narrowed down to an evaluation matrix comparing seven options, ranging from on-campus self-hosting to fully automated third-party administration. One intriguing option sat somewhere in the middle of this list: purchasing "infrastructure-as-a-service" through a cloud computing provider.

Current trends suggest that colleges and universities are incorporating elements of contemporary cloud hosting into their IT operations, often through the Amazon AWS service. Instead of managing their own virtual machine server farms, technology departments can issue custom operating systems on demand, complete with preinstalled applications, and migrate or decommission them just as quickly. Outsourcing technologies to domain specialists, such as the Digital Commons institutional repository platform, offers the advantage of a large infrastructure supporting such value-added features as customized readership reports and journal production tools. Disadvantages often include the expense attached to these services, and in the case of Digital Commons, ethical considerations. Bepress, the Digital Commons parent company, was founded at the University of California, Berkeley, but subsequently acquired by Elsevier, an academic publisher with a commercial interest in limiting open access publishing ("Elsevier Acquires Bepress" 2017). Partial outsourcing, in the form of self-managed infrastructure, features its own challenges, chief among them the need to maintain in-house technical skills in the form of DevOps, a contemporary technology practice integrating elements of software development with systems administration.

As one of the most successful cloud computing providers, Amazon supports corporate clients with requirements far beyond the scale of a small academic library, but the market is highly dynamic, with smaller competitors offering affordable packages for both individual and small business clients. After conducting interviews with marketing and technical representatives from several companies, the Cheng Library retained dedicated cloud hosting services using a regionally headquartered provider with several globally distributed data centers. Once arranged, a virtual machine was provisioned to our exact specifications, with externally managed block storage attached to the host.

Although not as flexible as Amazon cloud storage, in which data volumes are represented as objects independent of file systems, block storage offers a simple budgeting arrangement under predictable terms. Our base environment was recreating using a supported version of the GNU/Linux operating system, due to its Free and Open Source licensing, wide developer support (including the majority of the DSpace developer community), and the availability of all required software packages within the main distribution repository. With the exception of DSpace itself, every web server, Java application, and database management system is readily obtainable. Network bandwidth and CPU processing are billed at a flat rate, with usage monitored through an online console, seen in the screenshot below.

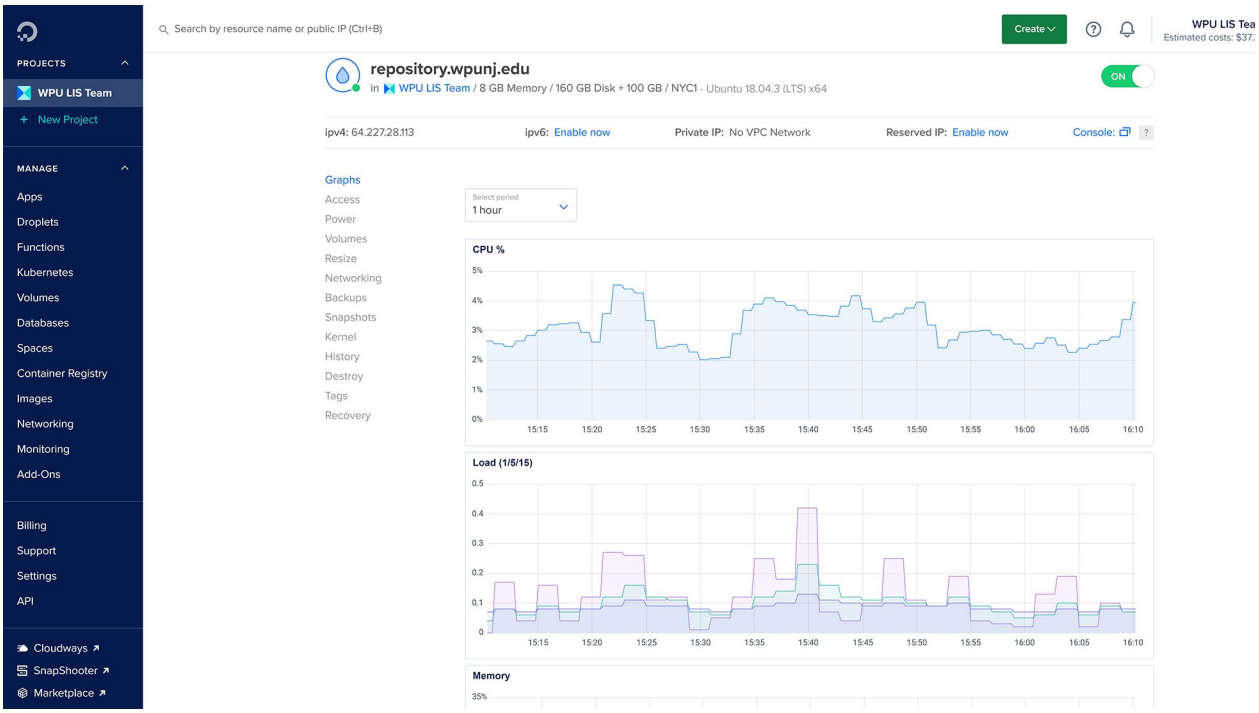

Image 1: Online management console.

An additional consequence of the DSpace application's portability is the insight required when encountering subtle differences between application dependencies. Building DSpace on UNIX sometimes requires distinct software versions and configurations when compared to a Microsoft Windows environment. However, the DSpace community, and open source developers in general, are an excellent source of information and documentation, and the software directly provides detailed debugging and error messages in the form of timestamped log entries. For veteran researchers, particularly librarians, uncovering bugs and troubleshooting performance issues can easily become second nature, but with careful planning such issues are infrequent.

After transferring the repository domain name to our new host and integrating the university's authentication service, over 2,000 archival packages representing established collections and assets were ingested into the new environment. The migration was a success, with no loss of content or persistent identifiers, and completely transparent to our user community. Technology, of course, evolves constantly, largely in the form of security updates and operating system upgrades. Moving forward, we hope our accomplishment and the lessons learned will lead to further collaboration with our university IT department, so that we can collectively participate in the next upgrade cycle. In the meantime our cloud hosting environment provides additional opportunities for creating, testing, and developing future projects and initiatives. As the technology becomes more ubiquitous and affordable, incorporating cloud computing into a library's technology portfolio will undoubtedly become a popular means for enabling greater participation in academic technology.

WORKS CITED

"Elsevier Acquires Bepress, a Leading Service Provider Used by Academic Institutions to Showcase Their Research." 2017. Elsevier. August 2, 2017. https://www.elsevier.com/about/press-releases/corporate/elsevier-acquires-bepress,-a-leading-service-provider-used-by-academic-institutions-to-showcase-their-research.