NJLA College and University Section
and
ACRL New Jersey Chapter

Digital Institutional Repository Management at William Paterson University: Planning and Staging a Seamless Migration

http://cus.njla.org/content/newsletter/Fall2020/digitalrepository

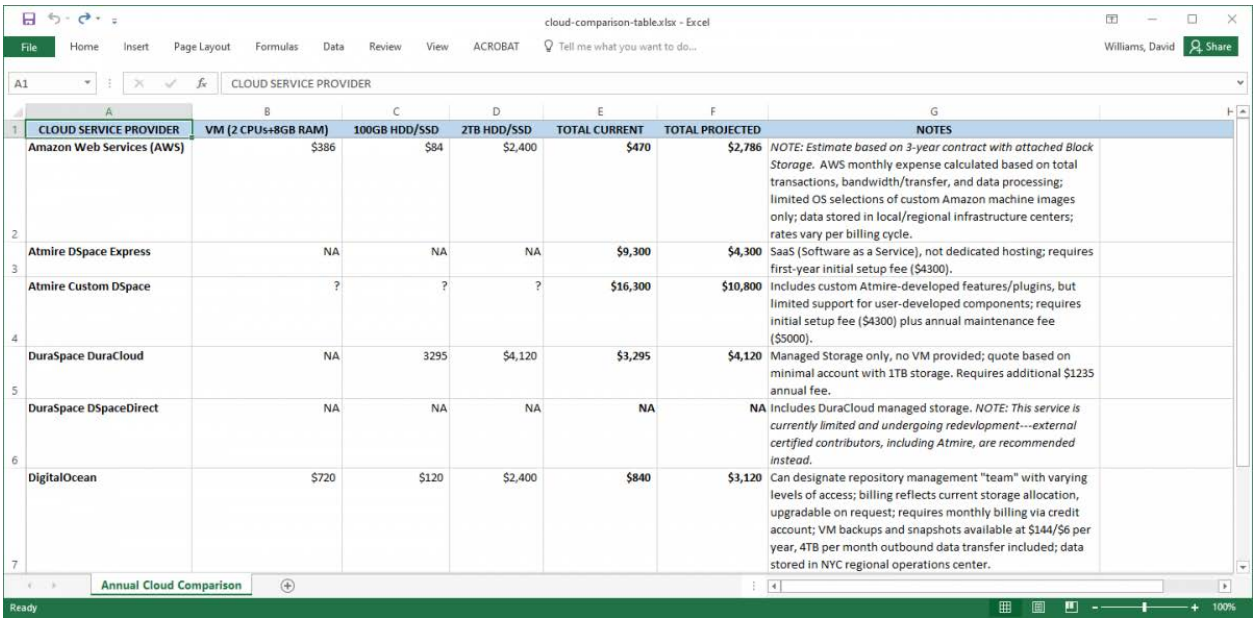
By David J. Williams, MA, MLS

Launched in early 2018, WPSphere is the David and Lorraine Cheng Library Digital Institutional Repository, charged with collecting, preserving, and disseminating the intellectual output of William Paterson University. As a replacement for the institution's previous CONTENTdm digital collection management service, WPSphere also houses a series of digitized historical materials, including scanned photographs, yearbooks, and student newspapers.

Created using DSpace, an open source digital repository system, the service can be installed in a wide range of operating environments. One benefit of this portability is the opportunity to apply a variety of supporting technologies and resources. If these resources are circumstantially limited or otherwise suboptimal, however, issues can occur; as with any complex technology project, careful planning is essential. Fortunately, an additional benefit of portability is that well-designed digital repositories support and enforce archival practices, ensuring that the preservation and maintenance of ingested collections is "baked in" to the system. That said, an expression common among Information Technology professionals states, "you can't really say you have a backup until it's been successfully restored."

An assessment conducted in September of 2019 revealed technical issues capable of rendering WPSphere unreliable. The existing collections were stable and accessible, but expanding usage to support future university initiatives was problematic. In the larger context, an Institutional Repository, digital or otherwise, embodies archival principles, chief among them authority. If the structures in place for preserving and describing collection materials are not trustworthy, they must be reconsidered.

Requirements were gathered and technical specifications analyzed, revealing the need for a flexible storage architecture within a stable hosting framework. The results informed the development of an evaluation matrix (see chart below), comparing a wide range of potential solutions ranging from outsourcing technical administration using specialists in the digital repository field to working within the available campus information technology domain. Narrowing down the list to seven options, in-person and remote interviews were conducted with a series of technical and marketing representatives.



cloud-comparison-table.xlsx - Excel

CLOUD SERVICE PROVIDER	VM (2 CPUs+8GB RAM)	100GB HDD/SSD	2TB HDD/SSD	TOTAL CURRENT	TOTAL PROJECTED	NOTES
Amazon Web Services (AWS)	$386	$84	$2,400	$470	$2,786	*NOTE: Estimate based on 3-year contract with attached Block Storage.* AWS monthly expense calculated based on total transactions, bandwidth/transfer, and data processing; limited OS selections of custom Amazon machine images only; data stored in local/regional infrastructure centers; rates vary per billing cycle.
Atmire DSpace Express	NA	NA	NA	$9,300	$4,300	SaaS (Software as a Service), not dedicated hosting; requires first-year initial setup fee ($4300).
Atmire Custom DSpace	?	?	?	$16,300	$10,800	Includes custom Atmire-developed features/plugins, but limited support for user-developed components; requires initial setup fee ($4300) plus annual maintenance fee ($5000).
DuraSpace DuraCloud	NA	3295	$4,120	$3,295	$4,120	Managed Storage only, no VM provided; quote based on minimal account with 1TB storage. Requires additional $1235 annual fee.
DuraSpace DSpaceDirect	NA	NA	NA	NA	NA	Includes DuraCloud managed storage. *NOTE: This service is currently limited and undergoing redevlopment---external certified contributors, including Atmire, are recommended instead.*
DigitalOcean	$720	$120	$2,400	$840	$3,120	Can designate repository management "team" with varying levels of access; billing reflects current storage allocation, upgradable on request; requires monthly billing via credit account; VM backups and snapshots available at $144/$6 per year, 4TB per month outbound data transfer included; data stored in NYC regional operations center.

(Above)*: Evaluation matrix.*

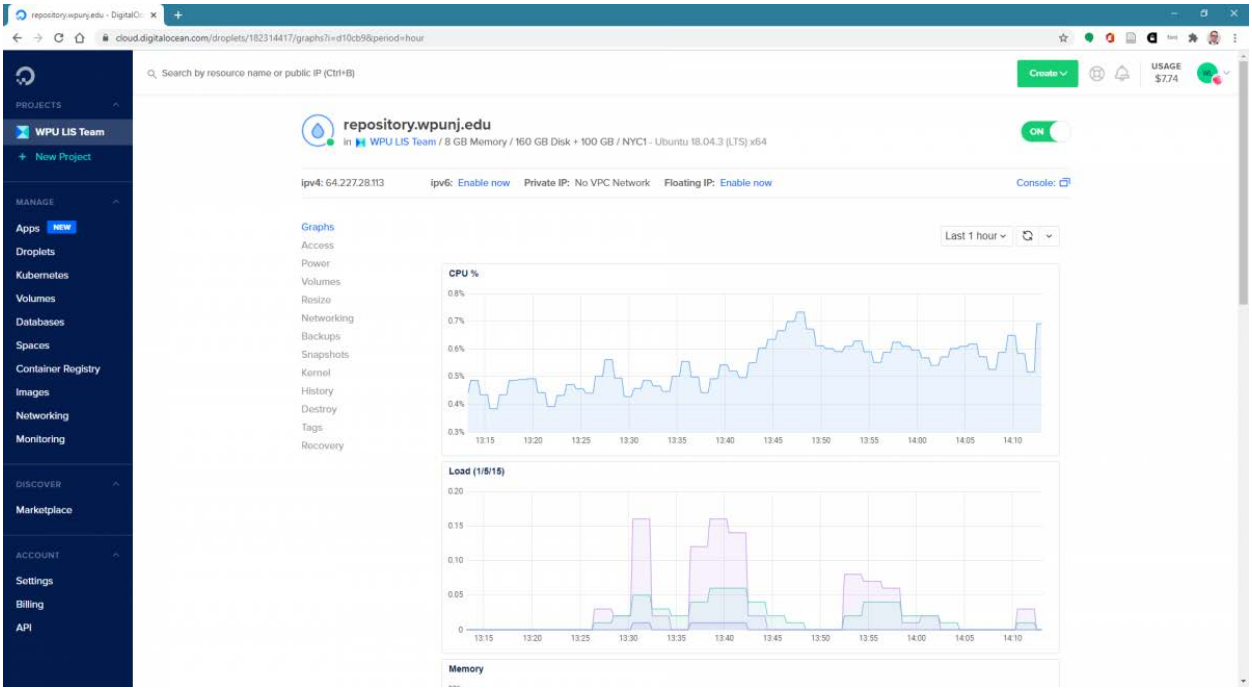
The earlier decision in favor of self-hosting was partially motivated by the expense of maintaining and expanding CONTENTdm, but for organizations interested in "insourcing," self-assessment is in order. Support available from IT offices varies from campus to campus, due in large part to the evolving nature of technology operations in higher education. The traditional approach, common to a variety of organizations, focuses on the hardware and infrastructure necessary for general business operations, with an emphasis on security. As demand for online learning and digital scholarship increases, separate academic technology units arise, embodying a service-oriented approach through directly engaging stakeholders and research communities.

Current trends suggest colleges and universities are incorporating elements of contemporary cloud hosting infrastructure into their IT operations, either externally, using Amazon's AWS services, or internally (as with the CUNY Graduate Center's Academic Commons). Outsourcing specific technologies using domain specialists, such as the archival content management services provided by Digital Commons, offers the advantage of a large development infrastructure supporting such value-added features as customizable reporting and electronic journal production tools. Disadvantages include the expense attached to these services, and, in some instances, ethical considerations. Digital Commons, for example, is owned by Bepress, a company initially created at the University of California, Berkeley but recently acquired by Elsevier, an academic publisher with a commercial interest in limiting open access publishing.

In the absence of outsourced resources, librarians need to assess their in-house technical skills, particularly in the area of Systems Administration. Openness is also crucial to a successful migration process: digital materials, packaging formats, and associated metadata provided using open standards enable introducing change without disrupting existing end-user services. This transparency is the ultimate goal of a migration: all URLs and references to collections and items should remain identical, every component element must transfer without loss, and the final result—the migration itself—should be effectively invisible.

After careful consideration, the library retained dedicated hosting services using DigitalOcean cloud infrastructure. A virtual machine was provisioned to exact specifications, and externally managed block storage, resizable on demand, attached to the host. Although not as flexible as

Amazon S3-style cloud storage, block storage offers a simple budgeting arrangement with predictable terms. Bandwidth and processing are billed at a flat rate, with usage monitored through an online console (see screenshot below). The base environment was recreated using Ubuntu's LTS (Long Term Support) version of GNU/Linux, due to its Free/Open Source licensing, wide developer support (including members of the DSpace community), and the availability of all of required software packages within the main distribution repository—no customization necessary.



(Above)*: Online management console.*

One additional consequence of DSpace portability is the insight required when subtle differences between versions of the application's dependencies are introduced. But the DSpace community, and open source developers in general, are excellent sources of information and documentation, with the software itself providing detailed messages by way of timestamped log entries. For veteran researchers like librarians, uncovering bugs and troubleshooting performance issues can easily become second nature.

After transferring the domain name to the new host, the university's Single Sign-On authentication service was carefully integrated and over 2000 archival packages ingested. Having completed the migration, we look forward to enhancing and further developing the repository as part of our efforts to support our research community. And, should the need arise to conduct another bottom-up migration, we can be confident in the stability and sustainability of our archival institutional assets.

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