Cloud-based Digital Preservation Services for Small or Midsized Institutions: Results of a pilot study of Archivematica + DuraCloud

Kevin C. Miller; Pepperdine University; Malibu, California

Abstract

This paper presents the results of a four-month testing period with ArchivesDirect, the hosted Archivematica + DuraCloud digital preservation service launched in February 2015. A discussion of the challenges that small or midsized institutions face when preserving their digital assets is followed by a contextualization of the new platform within the evolving ecosystem of web-based digital preservation services. The paper closes with a candid and critical analysis of the ArchivesDirect service itself as it stands at initial release. This presentation aims to contribute to the ongoing discussion of digital preservation challenges at under-resourced institutions while providing new and critical information on the web-based solutions emerging to serve this community.

Introduction

In August 2014, the Digital POWRR group published an important white paper titled “From Theory to Action: ‘Good Enough’ Digital Preservation Solutions for Under-Resourced Cultural Heritage Institutions” [1]. The paper clearly articulates the fiscal challenges that libraries, archives, and museums face in rising to meet the standards and best practices of digital preservation, and provides a candid environmental scan of the open-source and vendor-based services available for preserving digital assets. At Pepperdine University—like many other midsized institutions—we recognize the challenges detailed by the white paper, and have entered the fray, searching for the solution that best fits our digital preservation needs. While standards and best practices for digital preservation have coalesced, affordable solutions for their implementation are still emerging, and the field is fast moving. The principal challenge for small or midsized institutions faced with digital preservation, therefore, remains getting from theory to practice.

This paper presents the results of Pepperdine University’s pilot test of ArchivesDirect, a hosted digital preservation service that launched, subsequently, in February 2015. Pepperdine University was one of nine institutions to participate in the pilot program. A joint venture of Artefactual and DuraSpace, the new service seeks to leverage the open source micro-services of Archivematica—for ingest and processing—in a bundled, hosted instance that combines seamlessly with DuraCloud—for storage and maintenance. As the POWRR group white paper and other recent surveys [2] have pointed out, a single platform solution for all of an institution’s digital preservation needs remains rare. By pairing a hosted Archivematica instance with DuraCloud, Artefactual and DuraSpace hope to introduce a one-stop digital preservation solution to the market that can compete with vendors like Preservica (Tessella) and Rosetta (Ex Libris).

After discussing the challenges that small or midsized institutions face when preserving their digital assets, I will situate the ArchivesDirect platform within the evolving ecosystem of web-based digital preservation services. My analysis of the platform itself, as it stood at the close of the testing period, will follow, based on criteria identified by *Trustworthy Repositories Audit & Certification: Criteria and Checklist* [3] and the POWRR Group white paper. This new offering is unique in that it is based on *open source* micro-services and software solutions. But will ArchivesDirect satisfy the workflow requirements of institutions with limited resources and FTE devoted to digital preservation activities? How does the actual product perform? And how much are institutions willing to pay for streamlined access to these services and the technical support and peace of mind that comes with a vendor solution?

Digital Preservation Challenges Facing Small and Midsized Institutions

No matter its size, a repository mandated with the task of preserving digital assets must be prepared to scale its preservation activities to the *needs* and *means* of its contributing institution or defined community [4]. Digital preservation, defined here as “the series of managed activities necessary to ensure continued access to digital materials for as long as necessary” [5], therefore, rests on the administrative ability of an institution to support its needs with adequate means. Small or midsized institutions, which, in the case of academic institutions, are defined here as having less than 5,000 students and between 5,000 and 15,000 students respectively, may feel this tension with particular acuteness. Serving a student population of approximately 7,500, Pepperdine University Libraries are well familiar with this tension and the several challenges that face small and midsized institutions tasked with digital preservation.

According to a recent survey conducted by The Bishoff Group, academic libraries cite a lack of funding, other priorities, a lack of expertise, a lack of administrative support, and not knowing where to start as the primary roadblocks to implementing a digital preservation program [6]. Interestingly, this survey targeted 145 academic libraries from non-Association of Research Libraries (ARL) institutions, suggesting that the response group primarily represented the experiences of small and midsized institutions. I should also note that DuraSpace, one of the partners behind ArchivesDirect, commissioned the survey. Specifically, 73% of respondents cited a lack of funding as the primary element preventing the launch of a digital preservation program. The impact of fiscal restraints is noted by an earlier survey of institutions faced with preserving electronic scholarly literature, which cites the top three concerns as additional costs, lack of staff resources, and budget concerns [7]. A third study, conducted in 2013, surveyed a variety of academic libraries and concluded that most institutions have no direct funding for digital preservation, drawing funds instead from existing budgets for IT, collections, archives, or digital initiatives [8]. The same study found that the most common staffing for digital preservation activities was 1 Fulltime Equivalent (FTE), although many reported less, and most individuals doing digital preservation work had other primary duties as well. On this topic, a 2013 NDSA survey on staffing for digital preservation found that most organizations would prefer to have nearly twice as many FTEs as they currently had working on digital preservation activities [9].

As indicated by The Bishoff Group study, a second inhibitor to digital preservation activities is the shifting of resources to other priorities. At Pepperdine University Libraries, for example, our digital initiatives program initially focused on digitizing for dissemination, access, and research. Although we were careful to follow best practices in terms of generating archival quality digital images and documents, the emphasis was on curating digital collections for the research community rather than developing a long-term preservation program. The widespread use of platforms like CONTENTdm, DSpace, Fedora, and Digital Commons, and the slower emergence of more robust digital preservation vendor options suggest that the prioritization of access over preservation is a trend among academic institutions.

Related to the issue of staffing, small and midsized institutions also grapple with a lack of technical expertise or practical knowledge of digital preservation issues and activities. Academic libraries with relatively small staffs are often daunted by the complexities of digital preservation issues and the high bar of technical knowledge required to fully understand and implement best practices. This lack of training among staff in such organizations may also be compounded by "change fatigue," in which a history of organizational and technological change at an institution leads staff to greet digital preservation initiatives skeptically as an additional burden to their already heavy workloads [10]. The result is a lack of engagement among staff and a feeling of disempowerment or paralysis when confronted with digital preservation issues.

Although the challenges are many, there are solutions—some well tested, some emerging. Regarding education and training, information schools are increasingly making digital curation and digital preservation a part of their curricula, and continuing educational opportunities are increasing for professionals already on the job. While continuing education for digital preservation remains exploratory and experimental, some robust, sustainable programs have emerged, such as the Digital Archives Specialist (DAS) program run by the Society of American Archivists [11]. Collaboration, both internally across departments and externally with other institutions, remains a means for some small or midsized institutions to overcome the challenges they face. Participation in collaborative e-journal preservation initiatives, like LOCKSS or Portico, is common, while others partner with massive digitization collectives like Hathi Trust or the Internet Archive. Still other collaborative options include regional consortia, state-based digital archives, and digital preservation networks, such as the LOCKSS-based MetaArchive Cooperative. Still, even as institutions grapple with the decision to build a digital preservation program based on open source software, a collaborative partnership, or a hosted vendor service, they can take immediate, incremental actions to start preserving their digital assets without delay. As articulated by the Digital POWRR white paper, "It is appropriate to focus efforts on the activities we *can* perform in the next six to twenty-four months to steward our digital content, rather than wait a decade for a potential perfect solution" [1].

The Rise of Cloud-based Vendor Solutions for Digital Preservation

According to The Bishoff Group survey, small and midsized institutions are turning to hosted vendor solutions for digital preservation in large numbers. Of the respondents with existing digital repositories, 78% indicated that they used hosted services, while 31% indicated that repositories were locally managed (with some respondents indicating a combination of the two) [6]. Historically, cloud-based digital preservation solutions continue an evolution of services that traces back decades. Beginning in the 1990s, a shift from off-line to online digital storage provided an environment in which data migration is easy and inherent. Concerns about the longevity of particular media or carriers spurred the advent of file or bit stream preservation in a distributed, online preservation environment. The open source software movement emerged as an essential locus of digital preservation activity, democratizing intellectual property issues and providing renderers for proprietary formats. Finally, large, secure online storage networks, such as Amazon S3, appeared on the market as storage costs stabilized and lessened.

In just the last five years, the number of open source and vendor hosted digital preservation tools has greatly proliferated. The collaborative COPTR registry currently lists no less than 394 [12]. Respondents to a 2013 survey on digital preservation practices listed numerous platforms and tools, including Archivematica, Fedora, LOCKSS, Islandora, Archive-It, DuraCloud, MetaArchive, BagIt, and BitCurator, among others. The survey also supports the consensus, common in recent years, that no single system on the market provides a complete, one-stop solution for digital preservation needs. Said one respondent, "No system is perfect right now. It's a matter of getting a good enough system." And another: "It would be good if we could make these different utilities more systematic. Right now every collection is its own case, and we need an overall solution" [8]. The result has been a piecemeal approach to digital preservation in which institutions stitch together multiple tools and platforms to achieve a comprehensive workflow.

It is only recently that robust, all-in-one digital preservation vendor services have appeared in the North American market with pricing scaled to the library, archives, and museum community. Designed to be compliant with the *Reference Model for an Open Archival Information System* (OAIS), these systems offer "out of the box" hosted solutions for digital preservation activities, including processing, preservation metadata input, storage, and maintenance [13]. For example, Rosetta by Ex Libris grew out of a collaboration with the National Library of New Zealand, and now has clients among other national libraries and educational institutions, including the Getty Research Institute and the University of Utah. However, Rosetta is often left out of platform reviews (for example [1] and [2]) geared towards small and midsized institutions, possibly because of the steep capital investment required to sign up. Another contender is Preservica by the United Kingdom-based Tessella, which has had a client list of national libraries and archives in Europe for well over a decade. Preservica, available in hosted or local instances, is beginning to attract a list of North American clients, including the Michigan State Archives, Colby-Sawyer College, and the Hagley Museum. It is in this relatively uncrowded field of comprehensive digital preservation services that ArchivesDirect hopes to compete.

ArchivesDirect, an Evaluation

ArchivesDirect is the result of a partnership between two prominent names in open source digital collections software, Artefactual, based in Vancouver, BC, and DuraSpace, based in Winchester, MA. The new offering builds on two existing products, Artefactual's Archivematica, for *processing* digital content for preservation, and DuraSpace's DuraCloud, for the long-term *storage and maintenance* of digital content. Archivematica provides a single interface for running digital content (Submission Information Packages—SIPs, in OAIS terminology) through upwards of 30 open source digital preservation micro-services, such as checksum identification, virus checks, file format identification, normalization, etc. The process allows for the ingest or input of preservation metadata (PREMIS), and outputs Dissemination Information Packages (DIPs) or Archival Information Packages (AIPs) for access and storage utilizing external systems. DuraCloud, meanwhile, is a hosted service that provides backup, syncing, and health checks on files stored across multiple cloud storage providers, such as Amazon S3 and Amazon Glacier. Packaged as ArchivesDirect, the product offers two new services: 1) a hosted instance of Archivematica, which heretofore had only been available as locally installed and maintained open source software, and 2) direct, supported integration between the Archivematica dashboard and DuraCloud "spaces."

ArchivesDirect was released for subscription in February 2015. The current annual price is $11,900 for 1 TB of storage, one training session, and six cumulative hours of support. Additional storage is priced at $1,000/TB. These prices are comparable to Preservica, the other major soup-to-nuts digital preservation service on the market.

The following evaluation is based on Pepperdine University's participation in a pilot test of ArchivesDirect during the months preceding its release. The pilot program was intended to test and improve the hosted service and determine its effectiveness. Observations here, therefore, should be understood to reflect the product at the close of the testing phase. For purposes of consistency, I based my evaluation of the product's functionality on the POWRR Tool Grid version 1, which offers 21 criteria in five categories (Ingest, Processing, Access, Storage, and Maintenance) based on the OAIS reference model [14]. Additional assessment parameters are based on *Trustworthy Repositories Audit & Certification: Criteria and Checklist* [3] and related publications, like Bernard Reilly's "Planning for Digital Preservation: 20 Questions for Providers of Digital Storage Services" [15]. During the testing period, there was no new "ArchivesDirect interface" per se, so the following is essentially a description of how the two existing products (Archivematica and DuraCloud) worked together.

The Positives

- ArchivesDirect met all 21 aspects of the POWWR group's Tool Grid (see Figure 1). Joined together, Archivematica and DuraCloud do, in fact, provide a one-stop, comprehensive digital preservation solution that is OIAS compliant.
- The Archivematica web-based user interface is visually very nice and easy to use. The administration configuration tools allow customized workflows with the ability to preconfigure various decision points as the ingest process runs through the various micro-services.
- The ability to add preservation metadata (PREMIS) to each SIP, including rights and Dublin Core metadata, is very clear and easy to use.
- Both the Archivematica and DuraCloud interfaces are clearly organized and transparent in terms of the preservation actions taking place. Both systems provide clear reporting on outcomes and services, including—in the case of DuraCloud—some data visualization. Both systems provide clear documentation on systems, procedures, and policies.
- The Archival Storage interface provided by Archivematica provides a helpful organizational link to the AIPs stored in the relatively flat storage space of DuraSpace, which, on its own, is less easy to search and browse.

The Less Positives

- ArchivesDirect requires that source files be placed into a directory structure that includes (at a minimum) three hierarchical levels: a sync folder, a transfer location, and a transfer source (i.e., three nested folders). This is an infrastructural requirement of the system, which needs to "see" at least two structural levels within the target folder. This directory structure may not match the existing arrangement of an institution's digital content, requiring reorganization prior to ingest into ArchivesDirect.
- Ingest into ArchivesDirect requires a two-part process. First, local digital content is synced to a temporary transfer "space" within DuraCloud, which is then configured as a transfer source for Archivematica. After the AIP is processed and deposited in its permanent "space" within DuraCloud, the content residing in the transfer space must be manually deleted. This step, which is required by the Archivematica workflow, is not too burdensome, but it does make the process feel less streamlined.
- The ability of Archivematica to work with large files (greater than 1 GB) remains something of an unknown at the close of the pilot test. Although this was a focus of the latter stage of the pilot period, I cannot say conclusively that issues with processing large files have been resolved.

- There was some slowness in initiating transfers on the Archivematica side, although this will likely improve with future releases.

May Cut Both Ways

- The fact that ArchivesDirect is based on open source software and open data standards is certainly a positive. DuraSpace is a not-for-profit company and both companies are leaders in open source initiatives and members of the library, archives, and museum community. Furthermore, the fact that Archivematica's software is open source means that the client can "go local" at any time, reducing the feeling of vendor lock-in.
- Having said that, from a client's point of view, ArchivesDirect maintains a bit of an "open source feel" around the edges. While some users may appreciate the option to operate certain functions autonomously through command line tools or access to the pipeline configurations, many less technologically savvy users may feel out of their depth. Client satisfaction here will depend on the degree to which ArchivesDirect, as a hosted service, takes care of these details for clients in a seamless manner behind the scenes.

Conclusions and Recommendations

The authors of the POWRR Group white paper are not ashamed to admit that the majority of the five small to midsized universities that they represent have yet to reach Level 1 in some categories of the NDSA Levels of Digital Preservation matrix (protect your data) [16]. Achieving all of the benchmarks of Level 1 has been a challenge for Pepperdine University Libraries as well, and we still have few gaps to overcome. But, following the advice of the POWRR Group, there are steps institutions can take right now towards a digital preservation program: create an inventory of digital assets, work on a digital preservation policy, identify potential partners for support or collaboration, or sign up for a trial account with a digital preservation vendor service. All of these actions can be taken while developing and making the case for investment from higher administration.

With its roots in the open source movement, ArchivesDirect is a welcomed entry into the hosted digital preservation arena. It offers an OAIS compliant suite of processing services bundled with redundant, distributed storage and maintenance. The system is, for the most part, transparent and intuitively designed. It is likely that most of the infrastructural quirks detailed above will be smoothed out or circumvented in future releases; these are, after all, early days in the union of these services. The question that remains is one of affordability. Under-resourced institutions may well ask whether the price tag, which is comparable to Preservica, is worth it. Given that the processing side of the service is based on open source software that is freely available, the pricing of the hosted version sheds a stark light on the "free like a puppy, not free like an ice cream cone" characterization of open source software. Artefactual recommends that users installing Archivematica locally, which requires running a virtual machine in an Ubuntu (Linux) operating system, have dedicated IT staff familiar with Linux/Unix systems. As with most hosted services, ArchivesDirect liberates clients from dealing with server installation, maintenance, patches, and upgrades, covers compute and hardware costs, and provides training and expert support. Locally installed or vendor hosted, digital preservation comes with a dollar figure. No matter the choice we make, we must be advocates for the digital assets under our custodianship, work to raise awareness of digital preservation best practices, and convince higher administration that the value of digital preservation, done right, is worth the investment.

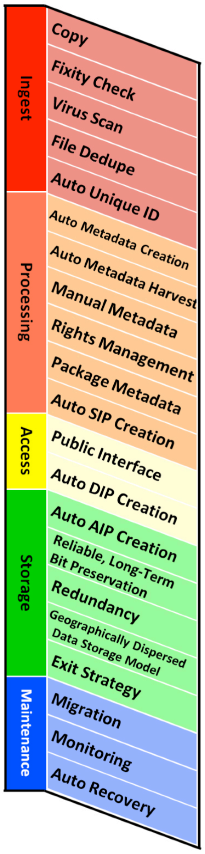

***Figure 1**. POWRR Tool Grid v1*

References

[1] J. Schumacher, L. M. Thomas, D. VandeCreek, et al., From Theory to Action: "Good Enough" Digital Preservation Solutions for Under-Resourced Cultural Heritage Institutions, IMLS White Paper, 1-23 (Aug 2014).

[2] M. G. Toussaint and S. Rounds, Report on Digital Preservation and Cloud Services, Minnesota Historical Society, 1-24 (April 2013).

[3] The Center for Research Libraries (CRL), Online Computer Library Center, Inc. (OCLC), Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC). Technical Report 1.0. (2007).

[4] The Center for Research Libraries (CRL), Archiving & Preservation: Ten Principles, www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/core-re (2007).

[5] Digital Preservation Coalition (DPC), Definitions and Concepts, http://www.dpconline.org/advice/preservationhandbook/introduction/definitions-and-concepts (2012).

[6] L. Bishoff & C. Smith, “Managing Digital Collections Survey Results,” D-lib Magazine, 21 (March 01, 2015).

[7] C. P. Meddings, “Digital Preservation: The Library Perspective.” The Serials Librarian, 60, 55-60 (January 01, 2011).

[8] M. B. Bergin. “Sabbatical Report: Summary of Survey Results on Digital Preservation Practices at 148 Institutions.” The SelectedWorks of Meghan Banach Bergin. Available at: http://works.bepress.com/meghan_banach/7 (2013).

[9] W. Atkins & National Digital Stewardship Alliance (U.S.), Staffing for effective digital preservation: An NDSA report: results of a survey of organizations preserving digital content (2013).

[10] A. K. Rinehart, P. A. Prud'homme, & A. R. Huot, “Overwhelmed to Action: Digital Preservation Challenges at the Under-resourced Institution. OCLC Systems and Services, 30, 1, 28-42 (January 01, 2014).

[11] H. R. Tibbo & Framing the Digital Curation Curriculum Conference, DigCurV 2013, View from Across the Pond: Opportunities, Gaps, and Challenges in Digital Curation Lifelong Learning. Ceur Workshop Proceedings, 1016 (2015).

[12] Community Owned digital Preservation tool Registry (COPTRR), http://coptr.digipres.org (2014).

[13] Consultative Committee for Space Data Systems, Reference Model for an Open Archival Information System (OAIS). Washington, D.C.: CCSDS Secretariat (2002).

[14] Preserving (Digital) Objects with Restricted Resources, Tool Grid, http://digitalpowrr.niu.edu/tool-grid/ (2013).

[15] B. Reilly, Planning for Digital Preservation: 20 Questions for Providers of Digital Storage Services, Northwest Document Conservation Center and Center for Research Libraries, https://www.nedcc.org/assets/media/documents/QuestionstoAskProvidersofDigitalStoragefinal.pdf (2007).

[16] M. Phillips, J. Bailey, A. Goethals, and T. Owens, “The NDSA Levels of Digital Preservation: An Explanation and Uses,” National Digital Stewardship Alliance, 1-7 (2013).

Author Biography

Kevin C. Miller holds a PhD in ethnomusicology and a Masters degree in library and information studies, both from the University of California, Los Angeles. Additionally, he received the Digital Archives Specialist (DAS) certification from the Society of American Archivists in 2014. He is currently Librarian for Digital Curation and Publication at Pepperdine University in Malibu, CA, where he also occasionally serves as Adjunct Professor in the music department. His interests include digital collections, institutional repositories, digital publishing, and digital preservation.