Environmental Scan of Distributed Digital Preservation Services: A Collective Case Study

Nathan Hall
Virginia Tech
560 Drillfield Drive
Blacksburg, VA 24060
1-540-553-2845
nfhall@vt.edu

M. Boock
Oregon State University
121 The Valley Library
Corvallis, OR 97331
1-541-737-9155
michael.boock@oregonstate.edu

ABSTRACT

This paper uses a collective case study to reveal similarities and differences in the operations and service models of nine distributed digital preservation services. The study uncovers a wide range of organizations and technical variations among the nine services, but finds that they can be grouped into three basic service models.

KEYWORDS

Digital preservation service models, digital libraries, digital curation, case study

1 INTRODUCTION

During the 2016 Annual Steering Committee Meeting of the MetaArchive Cooperative, the Steering Committee initiated an environmental scan of services provided by comparable digital preservation consortia and vendors. This study is a high-level overview of digital preservation service models, and includes an exploration of organizational aspects such as governance, support and training, documentation, community facilitation, outreach and communication, marketing, and membership; as well as technical aspects such as functionality, setup and configuration, content, ingest, storage, security, access, and integration; and finally, compliance with selected National Digital Stewardship Alliance Levels of Preservation[1].

Digital formats are highly sensitive to obsolescence, corruption, and degradation, elevating the importance of strategies and infrastructure for preserving digital records. Successful digital preservation “combines policies, strategies and actions” that ensure digital content survives in a usable form with an overarching goal to safeguard public records, scientific advances, and cultural heritage[2]. Many libraries engaged in digital preservation are “increasingly committing to the use of large-scale, comprehensive, distributed digital preservation systems” [3][4]. Such systems help libraries preserve greater amounts of data efficiently and cost-effectively according to standards and best practices.

A number of programs and providers offer similar digital preservation services, many of which are backed by the same third party vendors. However, they differ in their organizational models, strategies, and sometimes, architectures. The increasing number of options allows libraries to choose among several categories of service provider. Libraries might make financial and technical commitments based on the stated values of their service providers and peers within their community, how rigorously a service provider follows established standards, or based on the cost of the service and how it fits in the library’s budget and strategic plan. While we did not undertake a cost comparison since that information would be proprietary and/or negotiable, this study compares organizational and technical aspects of large-scale digital preservation services so that libraries can be better informed when deciding which providers make the most sense for them.

2 METHODS

Through documentary analysis, surveys, and interviews[5], we reviewed operations and service models of nine (MetaArchive, APTrust, DPN, TDL, DuraCloud, Preservica, Chronopolis, Rosetta, and Arkivum) distributed digital preservation service providers. We analyze governance, organizational structure, support and training, documentation, community facilitation, outreach and communication, marketing, membership, compliance with NDSA Levels of Preservation, functionality, setup and configuration, content, ingest, storage, security, access, and integration. We conducted a collective study[6] because these programs provide comparable functions yet have different organizational models and our purpose is to “reveal the properties of the class” of programs[7]. This analysis is a collective case study because it is an exploration of multiple programs in order to investigate distributed digital preservation services[8]. Schultz and Skinner[9] conducted a similar study that compared underlying technologies for three distributed digital preservation systems: Chronopolis, University of North Texas, and MetaArchive.

We created a rubric to compare each program across the categories listed above. Some of the information to complete this rubric comes from the organizations’ websites, and we received additional information through personal communication with executive directors or other program officials. Because we spoke with the representatives of these organizations about facts about their respective organizations, rather than opinions, Virginia

Tech's Institutional Review Board waived the review requirement. We compared the profiles of the nine service providers and analyzed similarities and differences within the sample along the lines delineated by the question categories. The themes and the comparative analysis together describe how the nine programs differ in their approach to managing digital preservation services.

We identify several limitations to this study, some of which are general to case study research[8][10][11]. Additionally, due to the informal nature of the semi-structured interviews, some participants may have interpreted questions differently from each other, or the follow-up questions may have gone in different directions in each interview. As a result there is limited basis for comparison in some categories. The third limitation is that this is a fast-moving area and some of these programs are planning and initiating new services at the time of writing. Some findings therefore may be out of date by the time of publication.

3 RESULTS

This section organizes the findings into categories for comparing similarities and differences among the program in the sample.

3.1 Organizational Aspects

3.1.1 Governance

All organizations in the sample provide their users with digital preservation services, but each one articulates its **mission** in distinct ways. The mission statements published on their websites each note the importance of collaboration or connections with other preservation and access systems. The Arkivum and Preservica websites state the sectors they serve without focusing on the content. Other providers stated that their purpose is to ensure the longevity of cultural heritage digital content, but do not provide business model details in their mission statements. DPN and Preservica list types of digital resources they preserve while Chronopolis and TDL discuss the mission of preserving digital content more generally. APTrust, Chronopolis, and DuraCloud each identify a **values** statement, and DPN indicate that they have one in development. The other institutions in the sample do not have a formal values statement.

There are three distinct **organizational** types with equal distribution in the sample. 1) APTrust, TD, and Chronopolis are each legally constituted as a part of a university. 2) Preservica, Arkivum, and Rosetta are commercial services. 3) DPN, DuraCloud, and MetaArchive are non-profit limited liability companies. Leadership varies across the sample, although MetaArchive, DuraSpace, DPN, APTrust, Preservica, and TDL are governed by boards or steering committees. MetaArchive, APTrust, DPN, and TDL all accomplish technical and non-technical development work through committees. Rosetta and Preservica have active user groups that meet regularly online, and at regular meetings and conferences. Arkivum has an executive board which includes investors and senior management, but does not have a formal user group.

Table 1-Member Community[12] demonstrates opportunities for community engagement associated with each preservation service. Community discussion are monthly calls or online forums. In-person gatherings are annual or bi-annual meetings. Mailing lists distribute service and product updates.

3.1.2 Support and Training

Most of the services provide **training** (see **Table 2-Member Support and Training**[13]) in the form of new customer orientations, instructional webinars, workshops, and/or video tutorials. DPN identifies itself as more of a catalyst organization instead of an educational organization, so they partner with other programs such as APTrust and DuraSpace for user support, and with AVPreserve and Educopia for curriculum development. TDL, APTrust, Arkivum, and MetaArchive provide in-person orientation and training either remotely or on site. Arkivum, Preservica, TDL, DuraCloud, and MetaArchive all offer training via webinars, and DPN plans to offer them in the future. APTrust and TDL both provide informal in-person workshops on-site to new members, whereas MetaArchive and DuraCloud offer in-person workshops on an irregular, ad hoc basis. Preservica provides briefings and workshops targeted at specific users and at conferences. DuraCloud, Preservica, Arkivum, TDL, and Rosetta all offer video tutorials through YouTube or Adobe Connect. APTrust also has training videos on YouTube but notes that they are slightly dated. All services offer **support** and troubleshooting via email, phone, or both. DPN offers informal support with legal agreements in digital preservation, but relies on APTrust and DuraSpace for formal technical support for members.

3.1.3 Documentation

The availability of comparable documentation in the form of Frequently Asked Questions and technical specifications varies among the nine providers. Rosetta, TDL, DPN, DuraCloud and APTrust, and to some extent MetaArchive have open technical documentation. Chronopolis, Arkivum, and Preservica on the other hand do not make their technical documentation openly accessible but it is available to customers. Among the three vendors, only Rosetta makes its documentation available to non-subscribers, including their AIP data model, system integration models, and user guides. DuraCloud and APTrust provide technical information on an openly available wiki (DuraCloud) or web site (APTrust) that includes knowledge bases of common issues, release notes, and detailed information on features and services. DPN provides an FAQ with documentation and code on their openly available GitHub. In lieu of a knowledge base, Arkivum has an online collection of case studies and white papers. Preservica, Chronopolis, and TDL do not have an open knowledge base. TDL had one in the past and may develop another one. MetaArchive provides technical specifications for hardware on its website, but in-depth technical resources and a knowledge base are currently restricted to users.

3.1.4 Community Facilitation

Nearly all of the providers we spoke with have a community, user, or customer-focused staff position. This position is responsible for communicating with users about services, and for facilitating community relations, meetings, and events. All except Chronopolis and Arkivum provide mechanisms for community discussion and product updates, either via mailing lists, Google Groups, or subscriber user forums. Each organization in the sample has a program in place for facilitating conversations with and between users. MetaArchive hosts regular monthly community calls that provide an opportunity for regular communication between community members. DuraSpace publishes a quarterly newsletter that includes updates about DuraCloud and ArchivesDirect. MetaArchive, Rosetta, DPN, DuraSpace, and TDL hold annual user meetings. APTrust members meet twice per year. Preservica facilitates user group meetings at conferences, and Rosetta, in addition to hosting an annual user group meeting, hosts quarterly working group web meetings and meets with advisory groups as needed.

3.1.5 Outreach and Communication

Table 3-Outreach and Marketing[14] illustrates that all of the programs engage with their community via social media. Most use Twitter, but some also use LinkedIn, YouTube, or Facebook. The commercial providers and DPN also sponsor conferences, and exhibits as well as join their non-profit and academic colleagues as conference presenters and panelists. Arkivum attends sector-based conferences in medicine and other fields.

3.1.6 Marketing

Preservica, Arkivum, and Rosetta engage in marketing, though MetaArchive, TDL, and APTrust do not. DPN has published flyers and co-sponsored events such as PASIG and Digital Preservation 2016. DuraCloud has exhibited at conferences, but does not engage in marketing. Chronopolis is developing promotional materials, but has never exhibited at a conference.

3.1.7 Membership

There are slight variations in membership models and composition. MetaArchive has twenty-two members and over sixty participating institutions including consortia, academic libraries, public libraries, archives, and museums. Nearly two hundred institutions use DuraCloud, including cultural heritage and commercial enterprise users. DPN has over fifty members, including universities, consortia, and one commercial entity (Figshare). DPN anticipates future membership to include public libraries and cultural heritage institutions. APTrust has sixteen members, all of which are ARL member academic libraries, but is expanding membership to include public libraries and liberal arts colleges. Chronopolis is not a membership organization, but the replication nodes are at University of California at San Diego, National Center for Atmospheric Research, and the University of Maryland Institute for Advanced Computer Studies, and most users are at those institutions. TDL has twenty-two members drawn from higher education institutions, though they considering expanding their membership model. Arkivum serves approximately one hundred institutions, though some share a consortium account.

3.2 Technical Aspects

3.2.1 Setup and Configuration

MetaArchive members that host a storage server node follow specifications, instructions, and support to set up a server to connect to the MetaArchive network. Preservica provides a hosted Cloud Edition that requires no installation, as well as an Enterprise Edition that requires a local server and storage (e.g. Windows or Linux servers with Network Attached Storage) and a MySQL database for managing metadata. DuraCloud users set up an account and then utilize a web dashboard to manually ingest content, or they can install a DuraCloud Sync tool to automate content ingest. Chronopolis and TDL both use SSH or DuraCloud. APTrust uses an API. Arkivum sends hardware for the customer to install, and then Arkivum runs the software installation and configuration.

3.2.2 Content

Table 4-Content and Ingest[15] demonstrates that all programs in the sample are format agnostic in that they support all file formats, content types, metadata schemas, and structures. Furthermore, they all support large file sizes as well as BagIt bags. All of the services except for MetaArchive and ExLibris have a drag-and-drop ingest interface, though ExLibris and MetaArchive do have a simple graphic user interface for uploading files through a web browser.

3.2.3 Ingest

There is wide variability in ingest methods within the sample. MetaArchive has two ingest processes--one for public, live, web-based content, which is crawled and ingested via LOCKSS[16] plugins setup with parameters to automatically ingest repository content, and another for non-public content that members place on a web server with simple directory listings. APTrust partners bag their own content using the LOC BagIt specification, and then submit their bags through the APTrust API. Partners can track the bag's progress through final ingest into AWS. DuraCloud users can add content manually via drag-and-drop to a web dashboard, or they can automate ingest through the DuraCloud Sync Tool. DPN, Chronopolis, and TDL also use DuraCloud software to upload content. TDL users can also use a desktop client that syncs to the network from a folder, or they can use a command line method that pushes from a server to DuraCloud. Preservica users can upload content via a web dashboard, a local SIP creator, a networked transfer agent, or Preservica's Cloud Edition Bulk Upload Service. Arkivum customers move files onto a networked file share. Customers who use Arkivum's Perpetua product can also use AWS for ingest.

3.2.3 Storage

As indicated in **Table 5-Storage**[17] there is wide variation in how different services handle storage. MetaArchive creates seven copies at node sites. DuraCloud has multiple subscription plans that offer two-to-four copies on separate cloud storage providers. Preservica's Cloud Edition includes one copy in S3 or Glacier, while their Enterprise Edition includes options for local storage along with S3, Glacier, or Azure. DPN, Chronopolis, and Arkivum each use a three-copy model, though one of Arkivum's copies is stored in a secure offline environment by a contracted third-party vendor. APTrust's model has six copies. TDL uses S3 and Glacier, which together create copies at two locations.

Nearly all of the programs operate as hosted services so that members do not need to install, run, or maintain servers or other infrastructure. Three of the services (DuraCloud, APTrust, TDL) offer replication through Amazon Web Services Glacier and S3. Preservica and Arkivum both offer managed hosted services, as well as enterprise versions that run on local infrastructure. MetaArchive, as a cooperative, has unique requirements for its members, who must purchase and host servers to run LOCKSS software and replicate other members' content. This configuration reduces MetaArchive's membership fee by transferring IT costs to member staff time and hardware. Alternatively, members can pay a hosted storage service fee.

All services in the sample provide regular fixity checking. Arkivum has the most frequent intervals with monthly checks. TDL reported biannual checks. Most participants did not report fixity check frequency.

3.2.4 Security

DPN security is managed by contributing partners. APTrust uses data integrity and siloing through AWS to secure its data. APTrust has scheduled a security audit for the coming year. Preservica did not provide details about security beyond access roles and rights. All users must authenticate with a username and password, and each user account is set up with a series of roles which allowing users to see content based on their defined roles. These configurations can be as granular as file level permissions. Arkivum data uses file encryption and customer-supplied encryption keys. TDL and DuraCloud's content security employs cloud storage service provider protocols (e.g. Amazon, RackSpace). All DuraCloud content uses https for encryption and TDL uses SSH and requires login and access control settings. MetaArchive member server cache communications are SSL encrypted. The setup of MetaArchive server caches includes firewall and port settings to restrict access only to other network servers.

3.2.5 Access

Preservica, Arkivum, APTrust, Chronopolis, DuraCloud, and TDL all restrict access through authenticated user credentials. MetaArchive's access system differs in that each member institution is a storage node and provides its own server which is managed by a local systems administrator. Each node hosts a copy of data from other institutions on the network, but only designated system administrators have user accounts for member server caches. No member is permitted to access another member's stored collections. Login pages are only accessible to the host member institution and the MetaArchive central staff, and SSH is required for remote access.

Each preservation system has a separate method for restoring content. MetaArchive members submit a request to MetaArchive central staff who establish a secure connection to the nodes with copies of the requested content, and then the software constructs an uncompressed ZIP package and makes it available for the member to download. DuraCloud users utilize a retrieval tool to download content to a local environment. Arkivum customers copy files back out through a fileshare, or if the customers use Archivematica, they can download ingest packages from that interface. Preservica's Universal Access module allows users to search, find, and download publicly available content, though details are not available for bulk download and restoration processes. Preservica has many different levels of accessibility that can be implemented, in addition to user-defined restrictions. APTrust members restore content through the API, and TDL members use the DuraCloud interface, but TDL staff retrieve content from S3 and Glacier.

3.2.6 Integration

DuraCloud integrates with and backs up DSpace, DSpace Direct, Archive-It, Archivematica, DPN, and Chronopolis. Through the REST API, DuraCloud allows integration in Ruby and Java. Preservica supports automated workflows to bulk ingest exported DSpace, PastPerfect, and CONTENTdm data, as well as bulk ingest of SharePoint, Outlook and Gmail packages, and website harvesting data. Preservica also supports integration with catalog systems including ArchivesSpace, Axiell, CALM, and AdLib for data management. APTrust's PREMIS logs and API for bag ingest allow integration with any system. Arkivum integrates with CRIS, EPrints, DSpace, Pure, Archivematica, and Figshare. Chronopolis integrates with Archivematica via DuraCloud and DPN. MetaArchive integrates with DSpace via the LOCKSS plugin, or through the DSpace replication task suite.

4 CONCLUSIONS

4.1 Technical

Three main service models emerge in this study. Most of the services follow a consistent hosted service model. APTrust, Chronopolis, and TDL are all legally constituted within public universities and use third-party applications such as DuraCloud software or Amazon Web Services Glacier and S3. MetaArchive is unique in its distributed network of nodes, where each member/partner provides its own storage hardware and contributes more staff time. In this regard, MetaArchive operates more like a cooperative. DPN is different because it operates as an independent umbrella organization that links other programs together, adding further replication and geographic distribution.

4.2 Organizational

Prior to data collection we anticipated seeing significant differences between commercial and non-commercial providers, and these differences did manifest in organization and mission—some providers are privately held commercial firms and other providers have a non-profit model. The non-profit providers are either legally constituted within public universities or as independent LLCs. While mission statements and business models articulated by the commercial entities exhibit competitive business services around data preservation, the non-profit organizations used language expressing a direct interest in sustainability of the content itself for posterity's sake.

Several of the services in the sample expressed plans or an interest in expanding their services to public libraries. This expansion is good news for the preservation of cultural heritage materials and data in public libraries. Among the sample the opportunity to expand services to this population will create new opportunities and challenges in coming years.

4.3 Geographical

All of the surveyed programs except Arkivum are based in the United States. Arkivum and all of its data are stored in the UK-based data centers. Understandably, the scale of size and distance are different in the U.S. and U.K. so Arkivum's nodes are a few hundred miles apart as opposed to a thousand miles or more, as is the case with some storage nodes and third-party data centers in the U.S. In spite of this difference, Arkivum's data centers are still far enough apart to have different disaster threats. Furthermore, Arkivum is the only service in the sample that offers offline secure storage in its preservation model.

While we are not aware of recent studies that parallel the scope of this one, it is a fertile area for further research. Some participants in this study are preparing to announce new services at the time of writing so a follow up study would yield different findings. Additionally, a larger and more diverse sample would be useful. Eight of the nine participants are U.S.-based, and one is based in the U.K. We became aware of other service providers that we would include in a follow-up, yet even these are exclusively within English-speaking countries. Aside from the scope, a few other subjects warrant further investigation. An analysis of cost versus effort could serve institutions investing in distributed digital preservation. Spalenka[18] writes that administrators or digital preservation program managers "should not assume that an application bundling many digital curation and preservation functions together with a single user interface will necessarily provide an entirely comprehensive and worry-free experience" because some service providers assume that users have already addressed the basic preservation needs. A tool that identifies existing resources and gaps in digital preservation architecture and matches that assessment with digital preservation services that best address the gaps would be a useful application of these findings.

Hopefully this study assists decision makers identify the best digital preservation service providers for their organization, and expands our participants' understanding of the digital preservation service environment while also helping them identify strengths, weaknesses, opportunities, and threats of their respective programs and service models.

ACKNOWLEDGMENTS

We would like to thank Carly Dearborn (Purdue University), Sam Meister (MetaArchive), Mary Molinaro (DPN), Bradley Daigle (APTrust), Kristi Park and Ryan Steans (TDL), David Portman and Tom Hodge (Preservica), Sibyl Schaefer (Chronopolis), Matthew Addis (Arkivum), and Heather Greer Klein and Bill Bannon (DuraCloud) for their time, contributions, and cooperation in this study.

REFERENCES

[1] Phillips, M., Bailey, J., Goethals, A., & Owens, T. 2013. The NDSA levels of digital preservation: Explanation and uses. In Archiving Conference (Vol. 2013, No. 1, pp. 216-222). Society for Imaging Science and Technology. http://www.digitalpreservation.gov/documents/NDSA_Levels_Archiving_2013.pdf

[2] ALCTS Preservation and Reformatting Section, Working Group on Defining Digital Preservation. 2007. Definitions of digital preservation. American Library Association Annual Conference, Washington, D.C. http://www.ala.org/alcts/resources/preserv/defdigpres0408

[3] M. Boock & B. Davis. 2017. Next steps for building a flexible and robust digital preservation infrastructure at Oregon State University Libraries & Press. http://hdl.handle.net/1957/60365

[4] Hitchcock, S., Brody, T., Hey, J., & Carr, L. (2007). Digital preservation service provider models for institutional repositories: Towards distributed services. *DLib Magazine*, *13*(5/6). http://www.dlib.org/dlib/may07/hitchcock/05hitchcock.html

[5] http://ir.library.oregonstate.edu/xmlui/bitstream/handle/1957/61411/DDR_Environmental_Scan.pdf?sequence=1

[6] S. Merriam. 1998. *Qualitative Research and Case Study Applications in Education*. San Francisco, CA: Jossey-Bass, Inc.

[7] E. G. Guba & Y. S. Lincoln. 1981. *Effective Evaluation*. (p. 371) San Francisco, CA: Jossey-Bass, Inc.

[8] S. Merriam. 2009. *Qualitative Research: A Guide to Design and Implementation.* San Francisco, CA: Jossey-Bass, Inc.

[9] Schultz, M. and Skinner, K. (2014). Comparative Analysis of Distributed Digital Preservation (DDP) Systems. https://educopia.org/sites/educopia.org/files/deliverables/Comparative_Analysis_for_DDP_Frameworks.pdf

[10] S. A. McLeod. 2008. Case study method. Simply Psychology http://www.simplypsychology.org/case-study.html

[11] R. K. Yin. 2009. *Case Study Research: Design and Methods, 4th Ed*. Los Angeles, CA: Sage.

[12] http://ir.library.oregonstate.edu/xmlui/bitstream/handle/1957/61411/IPres2017_Table1_MemberCommunity.png?sequence=5

[13] http://ir.library.oregonstate.edu/xmlui/bitstream/handle/1957/61411/IPres2017_Table2_MemberSupportandTraining.png?sequence=6

[14] http://ir.library.oregonstate.edu/xmlui/bitstream/handle/1957/61411/IPres2017_Table3_OutreachandMarketing.png?sequence=7

[15] http://ir.library.oregonstate.edu/xmlui/bitstream/handle/1957/61411/IPres2017_Table4_ContentandIngest.png?sequence=8

[16] Reich, V., & Rosenthal, D. S. (2001). LOCKSS: A permanent web publishing and access system. *D-Lib Magazine*, *7*(6), 14. http://webdoc.sub.gwdg.de/edoc/aw/d-lib/dlib/june01/reich/06reich.html

[17] http://ir.library.oregonstate.edu/xmlui/bitstream/handle/1957/61411/IPres2017_Table5_Storage.png?sequence=9

[18] D. Spalenka. 2016. Some assembly required: Micro-services and digital preservation. Digital POWRR. http://digitalpowrr.niu.edu/some-assembly-required-micro-services-and-digital-preservation/