AFS | Vermont Folklife Center

Digital Preservation Storage Planning

Report

2017-06-12

Prepared By:

AVPreserve
253 36th Street, Suite C302 | Brooklyn NY 11232
634 West Main Street. Suite 202 | Madison WI 53703

bertram@avpreserve.com

May 30, 2017

Telephone
917.475.9630

Introduction

In spring 2017 AVPreserve (AVP) consultant Bertram Lyons was supported by the AFS Consultancy and Professional Development Program to work with the Vermont Folklife Center (VFC). The goals of the NEA-funded project focused on an assessment of VFC's digital preservation procedures and hardware, and the formulation of recommendations for improving their efforts and upgrading storage systems.

On April 10–11, 2017 Lyons traveled to Middlebury and met with Kolovos for the on-site portion of the consultancy. During this visit Lyons gathered data to assist VFC with both the NEA-funded aspects of his work and the AFS-supported project. Lyons conducted additional work off site over the course of the next month.

Current Status of Backup Storage

Current VFC digital storage hardware consists of an aging, RAID 10 configured, Dell PowerVault array with approximately 5 TB of usable storage distributed across two volumes. VFC initially deployed a Dell LTO3 tape library for local backup, but the hardware proved unreliable, and ultimately stopped working within 3 years of installation. As a bridge approach for backing up digital collections materials VFC maintains two sets of 5TB Western Digital external hard drives to which staff back up data on a monthly basis. One hard drive is stored on site in Middlebury and the other stored at the archivist's home 35 miles north in Essex Junction, VT.

Project Description

VFC staff are well aware of the limitations—if not outright dangers—of grounding their approach to file backup in use of commercially available, external hard drives. In an effort to improve their approaches, Kolovos sought out other options, in particular cloud-based providers of limited access, offline (or "dark") storage. These cloud-based services provided by Amazon, Oracle, and others offer a lower cost option for offsite backup, but Kolovos quickly found he needed additional support to navigate and evaluate the range of services offered, cost structures, and approaches to file ingest. As noted above, AVPreserve had already received an AFS Consultancy and Professional Development Program award to work with VFC and, in light of the existing arrangement, AVPreserve reached out to AFS for support to extend the scope of the consultancy. Lyons visited VFC in April 2017 to conduct his onsite review and remained in contact with Kolovos as he formulated his recommendations for the AFS-funded portion of the project.

Recommendations

Based on VFC's current state and priorities for the next three years, AVP determined that an offsite backup of VFC's holdings should be designed in concert with a digital preservation readiness review. Such a review is essential as a first step in planning for VFC's digital collections, and should precede (and inform) the acquisition of any digital preservation-specific software product. It provides a framework for building digital preservation services for the organization by addressing these core preservation foundations:

- Achieve Redundancy
- Organize Content Consistently
- Inventory All Content
- Checksums For All Content and Ongoing Health Checks
- Develop Ingest Protocols
- Develop Security And Disaster Recovery Procedures
- Budget for Backup Storage

We will detail these recommendations fully in our report. The remainder of this document is built around the assumption that the structured recommendations provided in the NEA report have been fully implemented prior to adopting the guidelines here discussed.

Backup Storage Overview

With digital content, a central component of preservation is duplication. Digital storage devices can fail entirely in the blink of an eye, unlike physical objects that will decay in poor storage over many years. Maintaining multiple copies of digital content on different types of storage devices in different geographic locations is the tried-and-true approach to ensure bit persistence. Organizations seeking to preserve digital content over time develop ongoing strategies to mitigate the risk of losing data to storage failures by having clearly documented schedules for backup routines and by ensuring multiple copies in multiple locations.

DIY or Paid Service?

As an organization evaluates the strategy for offsite backup storage for digital content, the first question to answer relates to the type of storage service that will be selected. Options include the purchase of an organization-owned storage device, e.g., a Network Attached Storage (NAS) device, that can be installed in another geographic location and synced to the main storage device (DIY-offsite); simple hosted online storage, e.g.,

Rackspace, Amazon Web Services, Microsoft Azure, Oracle Cloud (simple-cloud); or preservation service online storage, e.g., Duraspace, DPN, APTrust (managed-cloud).

For each case above (DIY-offsite, simple-cloud, and managed-cloud), an organization should weigh pros and cons based on their particular situation.

Decisions about what type of storage works best for an institution's backup needs should be influenced by such things as:

- The level of reliability or "uptime" required. *Do you need immediate access to your digital content or can there be delays of minutes or hours in retrieving it?*
- The number and types of users that need access to it.
- Types and amount of digital content. *How much storage do you need? At what rate will it grow?*

Cost is a major factor in deciding what type of backup storage to select for an organization. And, when determining cost, it is important to take all of the costs of managing digital storage into account. The total cost of ownership (TCO) considers all of the hardware, software, media, labor and overhead costs that go into installation, ongoing management, and even how you might move from one storage option to the next when the time comes to do so. Of all the costs, ongoing management is the highest, and so institutions are more and more looking to cloud storage as a means for alleviating the day-to-day costs and responsibilities for storing backups of their digital content.

Whether cloud storage is the answer for backup storage depends on a particular institution's local organizational structure, access to resources, and technology infrastructure.

> **Local Storage**: Storage offerings are as diverse as the institutions that they serve. They may be online only, or some combination of online, nearline, and offline. In all cases, there are associated costs to managing the servers and media on which digital content is stored. Staffing, facilities, and ongoing management of and upgrades to technology must all be factored into the costs of maintaining backup storage locally. It is valuable for digital collections managers to develop strong relationships with the IT staff that manage storage at their institution, so that they can work together to build the best backup storage environment possible for the digital content they wish to preserve over time.

Cloud Storage: Cloud storage is a service model in which digital content is maintained, managed, backed up remotely, and made available to users over the Internet. Examples of cloud storage include Amazon S3 and Glacier, and Google Cloud Storage. Many organizations are now using cloud storage for primary and for backup storage. This review focuses mainly on the use of cloud storage for backup storage.

Cloud providers offer different services, features, and performance levels based on costs and the intended market. A few considerations when assessing cloud storage options for backup storage include:

- **Latency**. How quickly does the system respond to requests for access to a digital file?
- **Geographic diversity**. Will your data be stored in one location or replicated up to multiple locations?
- **Security**. What services are in place to ensure your data is safe?
- **Disaster recovery**. What happens if the data goes away permanently?
- **Exit path policies**. How difficult is it to get your data out, either in chunks or as a whole?
- **Costs**. What are the costs to upload data into the cloud? What are the ongoing service costs? What does it cost to download your data or exit the service entirely?

AVPreserve has developed cloud storage vendor profiles[1] that offer an overview of comparative features across offsite storage providers. These profiles provide both a list of services that exist, as well as an overview of their services in the context of digital preservation storage.

Online, Nearline, or Offline?

One factor to keep in mind when evaluating the offsite backup storage strategy that will be best for your organization is whether you will need access to content immediately at the touch of a button, or whether you can stomach a small delay in retrieval of a few minutes or even a few days. All of the above approaches to storage can be configured to handle each of these service requirements.

Online, nearline, and offline are terms that are used to describe different types of storage architectures. These terms speak to the ease and immediacy with which data can be accessed, as well as the varying costs and scalability of storage.

[1] Available here: https://www.avpreserve.com/papers-and-presentations/cloud-storage-vendor-profiles/.

Online: Online in this context means that the data is immediately available to users on a storage system. Servers that host an institution's networked drives are examples of online storage systems. This is the fastest, but also the most costly of the three architectures—it is also the most predominant. Examples of online storage include flash and spinning disk.

Nearline: In this case, digital content is available to users with some lag time (from a few seconds to 60 seconds or longer). It is automated and networked in the same way that online storage is, but the media is different—typically a magnetic tape library. This tends to be an option for larger institutions with the resources to diversify their storage architectures.

Offline: Here, digital content is stored on a piece of media that requires a human to connect it to a computer to access the data on it. The most common of these media types used for digital preservation is magnetic tape. Offline storage is often used to back up digital content for long periods of time. It is cost effective, but also takes time to access because it is not connected to a network. For offsite backup storage, most organizations are willing to accept offline latencies of a day or two in order to achieve long-term cost savings. An assumption with offsite backup storage is that it is not functioning as a day-to-day access point for digital content, rather it functions as a disaster recovery option in the case that primary storage (or primary redundant storage) is ever damaged.

Why use one architecture over another for backup storage?

Cost. Offline storage tends to be the cheapest, but has drawbacks in that it is not actively managed by automated digital preservation processes like fixity monitoring.

Immediacy. Online storage has the quickest response time–typically content is immediately available when needed. It may not be necessary for all content to be accessed immediately. Second or third copies of digital files are often stored on nearline servers or offline LTO tape because they will not be accessed regularly and latency is higher.

Bandwidth constraints. Larger files, such as high resolution video, take time to transfer over networks. If quick access to high resolution files is required, the cloud might not be an ideal solution as it requires transfer over the internet, with available bandwidth in and out of your facility providing additional constraints.

Scale. The scale of digital collections can be massive. It may not be cost effective to store all digital content on online storage. As long as two copies are actively managed on online or nearline architectures, other copies can be stored offline (e.g., on LTO tape, or using a service such as Amazon Glacier), which tends to be the most cost effective method of storage.

Many institutions consider a hybrid on-premises and cloud solution for their overall storage architecture. Cloud storage vendors have options for both online, nearline, and offline storage that, together with local storage, may provide an institution with a redundant and secure approach to managing its digital content.

VFC's Approach

Having reviewed VFC's current infrastructure and practices, AVP recommends the best options for VFC's overall storage strategy currently include onsite primary storage networked with internet connections to a near-line cloud storage provider for secondary backups and an offline cloud storage service for disaster recovery backups.

Such a strategy may be articulated as follows:

- NAS - Local online storage (32 TB) - primary storage (includes preservation materials, access materials, and staging content)
- Amazon S3-lowAccess nearline storage - redundant of NAS - Local (only preservation materials)
- Amazon Glacier offsite storage - redundant of NAS - Local (preservation materials and staging content; could be substituted with Oracle Cloud or other dark storage providers)

Implementing Cloud Backup

Getting started with cloud-based storage (whether online, nearline, or offline) requires an initial review of pricing schemes for storage. Most services have separate costs for individual functions, e.g., upload, static storage, download, delete requests, access requests, geographic redundancy, or availability. Most services focus on grouping these costs into an average cost per gigabyte per month. Before creating an account with a particular storage service, it is advisable to compare these average costs and evaluate the general costs per month for downloads and uploads, too. These will help you get a clear picture of the total cost of ownership, as well as the cost for moving all content out of the storage environment if the need ever presented itself. Usually, downloading content is the most expensive part of cloud storage services. If you are using the

service as a disaster recovery offsite backup for your primary storage, then the likelihood you will download files is low. And, the time that you might actually download all of your content, will likely be a time when you are willing to pay the costs, since likely you would have had a disaster that caused you to lose all other copies of your content.

Additional concerns before selecting a provider of offsite cloud backup storage include whether or not the service has easy-to-use tools for loading content into the service, whether those be online interfaces, desktop applications, or command-line packages that you can employ to create automated or manual ingest processes. Additionally, many NAS operating systems now support marketplace applications that are designed to sync directly between your NAS operating system and particular service providers (e.g., AWS, or Azure).

Once you create an account with your preferred provider, you will want to adjust a few settings for your storage bucket, including geographic separation (will the bucket be replicated within the cloud service to a second geographic location?), level of access (will you use the offline storage buckets or the online storage buckets, e.g., AWS S3 vs. AWS Glacier?), and general configuration considerations, such as which areas of your primary storage will be backed up in this offsite storage environment, how frequently backups will run, or whether deleted files in the primary storage will also be deleted in offsite storage.

Disaster Recovery

On top of a backup strategy, digital preservation requires a plan for disaster recovery. How will the organization recover in the case of storage failures or hardware failures? There are two aspects that speak to the need in this case: having a plan, and having it documented in multiple places. This is necessary in order to ensure that sufficient backup and recovery capabilities are in place to facilitate continuing preservation of and access to systems and their content with limited disruption of services.

Budget for Backup Storage

For this project we prepared two cost scenarios, projecting each over the next 8 years (2017-2024). This allows one to see the cost of replacing hardware in 5 years.

Scenario 1:

- NAS - Local (32 TB) - main storage (includes preservation materials, access materials, and staging content)

- Amazon S3-lowAccess - redundant of NAS - Local (only preservation materials)
- Amazon Glacier - redundant of NAS - Local (preservation materials and staging content; could be substituted with Oracle Cloud)

Scenario 2:

- NAS - Local (32 TB) - main storage (includes preservation materials, access materials, and staging content)
- NAS - Offsite (32 TB) - redundant of NAS - Local (only preservation materials)
- Amazon Glacier - redundant of NAS - Local (preservation materials and staging content; could be substituted with Oracle Cloud)

By performing a directory and file analysis against all files currently residing in VFC storage, we were able to use current storage sizes as starting points for 2017. We also performed calculations against content per year to get an average GB growth per year (over a span of 2001-2016). We used that average to project growth moving forward per year and to show cumulative growth against current data. The following numbers represent a current state and general summary of VFC's two digital collections storage volumes (because duplication was found between the two volumes, the numbers represent a proposed "post-clean-up" scenario):

Media-processing volume (post-clean-up estimate)

- 40,698 files
- 1,7335 GBs
- Average growth per year (GBs) 2001-2016 = 108.32 GBs
- Average growth per year (%) 2001-2016 = 27.99%

Archive volume (post-clean-up estimate)

- 31,390 files
- 3,386 GBs
- Average growth per year (GBs) 2001-2016 = 205.05 GBs
- Average growth per year (%) 2001-2016 = 23.33%

To show future projections, using the above starting points, we gathered yearly storage costs by calculating per GB/per Year storage costs for Amazon S3-lowAccess, Amazon Glacier, and Oracle Cloud. We also generated a per GB/per Year storage cost for VFC's proposed NAS-local[2] based on a 5-year split of the storage capacity and its

[2] VFC is currently looking at a NAS recommended by their IT: SYNOLOGY RS2416RP+ (32TB).

overall cost. We added some ancillary storage media[3] every two years (replacement drives). We added a 15% cost of ownership on top of all yearly storage (this is a typical cost that appears and helps with overs and unders). And, we included VFC's current yearly IT costs (with a 10% rate of growth each year) for onsite technology services. From these variables, we are able to look at VFC's cumulative storage costs and yearly costs for the next 8 years[4].

Conclusion

VFC and AVP thank AFS for supporting this consultancy. The information provided to VFC will be used to inform practical action with our digital holdings going forward, and we hope this report is of use to other organizations and individuals struggling with similar challenges.

[3] Using this drive as an example: HGST 4TB Deskstar 7200 rpm SATA III 3.5" Internal NAS Drive Kit.
[4] See sample budget spreadsheets here: https://docs.google.com/spreadsheets/d/1NpGOL4V6RkZagxcntJA_qiSHdRGQtyV0HvJLgu4Oa5w/edit?usp=sharing.