AUDIOVISUAL
PRESERVATION
SOLUTIONS

Feet On The Ground:
A Practical Approach To The Cloud

Nine Things To Consider When Assessing Cloud Storage

by seth anderson
consultant
audiovisual preservation solutions
2014

AVPreserve | Media Archiving & Data Management Consultants
350 7th Avenue, Suite 1605 | New York, New York | 10001
917.475.9630 | www.avpreserve.com | info@avpreserve.com

Introduction

Planning and decision making for any archives and preservation project is fraught with consideration and re-consideration of every detail. Tasked with preserving and distributing valuable collections in an ever-changing landscape, determining which tools to leverage for these responsibilities is often overwhelming. Archives and archivists are only able to identify appropriate solutions for preservation and access by defining organizational and collection requirements in order to perform careful vetting of available services. This is especially true when considering emerging technologies, which vary widely in their functions and represent new requirements and potential risks. In the case of third-party services, such as cloud storage, the importance of the vetting process is evident. Ensuring a service can perform necessary functions and be trusted is the first step to realizing the potential resource saving and convenience these services provide.

When evaluating cloud storage providers, it is dangerous to assume such services are only storage and therefore uncomplicated or that requirements for storage are obvious and therefor inherently met by the service provider. Experience with any technology selection will prove the opposite. No two services are the same and the variance between services often represents the difference between successful implementation and a failed initiative. Never purchase a service without proper vetting; uninformed decisions risk loss of time, money, and even assets.

There is no all-in-one solution that will fulfill every archives' needs for preservation storage. Often, cloud storage services fulfill a portion of an organization's larger preservation infrastructure, providing secure back up for preservation copies or supporting delivery of access files from low-latency storage. **Vetting and selection is therefore the alignment of organizational and collection needs with the offerings and functionality of a service. This means defining your acceptance criteria for optimal functionality and understanding how a service will fit in your preservation environment.**

The difficulty lies in knowing where to begin and what questions to ask. The following nine topics provide a place to start. Use these to promote research and discussion, but also make sure the relevant considerations below are adequately documented in the service agreement as well.

1. Defining Requirements

Begin by deciding what and where third-party storage fits within your preservation environment. Cloud storage may be suitable to fulfill one or more aspects of an integrated digital preservation environment consisting of systems, storage, policies, and people. .

It is then necessary to understand exactly what role a third-party service will play, what services the service will not provide (e.g., SIP validation, data integrity validation, file characterization, among others), and how it may integrate with other technology prior to researching and vetting potential solutions. This narrows down the number of potential services to explore and ensures you do not waste your time exploring services that cannot fulfill your needs.

Frequently, organizations find value in building a multi-tier and distributed storage architecture to serve a wide range of issues encompassing financial considerations, user needs, redundancy, geographic separation and other functional and business requirements. There are many potential combinations of multi-tier storage architectures and various factors will drive decisions around the best one for a given organization. Typical best practice in archives is to maintain varying qualities of files for fulfilling preservation and access requirements. These have varying anticipated usage, security concerns and performance requirements, all of which dictate storage and bandwidth needs. For example, a video access copy, intended for browsing by a student at a university is best served by high-availability, low-latency online storage to provide accelerated access. This set of requirements may be well served by both on-premise solutions and third party cloud services. Selecting which one works best will come down to more granular technical and financial factors that are best identified through vetting against requirements.

Questions to ask:

- How often will materials be accessed and by which user types?
- How quickly do materials need to be retrieved?
- Are there particular specifications that need to be met for security, reliability, and uptime?
- What is the anticipated need for storage over the next 3 – 5 years?
- Are their existing policies about the type of storage required (e.g., tape vs. disk technology)?
- Are there existing policies about the geographic location of storage or budget expenditure (e.g., all servers or money spent must be within a specific state, region, country)
- What is the immediate, mid-term and long-term budget for storage architecture?
- Does the organization favor capital expenses over expenditures on services or vice versa?
- Does the organization have staff expertise to manage storage and preservation services?
- Are you using other cloud services that would make it advantageous to use cloud storage?

2. Vendor Assessment

The number of suppliers in the cloud services market continues to grow as cloud computing becomes more ubiquitous. Creating a short list of service providers can be challenging, even when requirements have been identified. A strong initial approach is an assessment of the vendor, not the service. Understanding the viability of a company, the maturity of its service and operations, and the satisfaction of its customers will provide a picture of the suitability of the service for your needs.

A good indicator of a vendor's maturity and viability is the composition, size, and satisfaction of its user base. Third-party vendors serve a number of different sectors, so familiarity with the unique needs of a preservation environment for archival materials will differ. It is not necessary that they understand your needs thoroughly, but it may improve your experience as a customer.

The size of client base can be a positive or negative factor. A service with a large, diverse user base can be evidence of a quality product or customer service, though such a large client group may affect the efficacy of customer service. Determining whether a vendor is a good fit with your organization can be tricky, but considering the potential working relationship is a necessary step in any technology selection process. It will set the stage for establishing trust with the selected service.

Questions to ask:

- How many years has the vendor been providing this service?
- Will the company be financially viable in five years? Ten years?
- What are the vendor's customer service protocols? What is the stated turnaround time for response to customer issues?
- How big is their customer base? What types of organizations and companies do they serve?
- Does the terminology and positioning demonstrate they understand the responsibilities and required functions of archives and preservation?
- How does the company offer support? Email only? Support plans for purchase?
- Are the terms and conditions, or terms of service, reasonable and acceptable?

3. Storage Hardware Management

When delving into the nuts and bolts of service evaluation, the focus should be the service's adherence (or flexibility) to basic digital preservation principles[1]. Getting into the weeds of storage hardware and technology likely exceeds most archivists' expertise, and distracts from the chief consideration at hand: **Can the caretaker perform the necessary preservation and access actions in collaboration with this service?**

With this in mind, when evaluating a service's hardware, the emphasis is not exclusively the type or architecture, but the service's management of its hardware. In support of preservation, the service should:

- Replicate materials in geographically separate locations,
- Monitor the health and age of storage hardware (estimated lifecycle of tapes and drives is 3-5 years), and
- Regularly update storage hardware and migrate data.

These are the core principles that arise in any digital preservation initiative, but are doubly important when assessing cloud storage options. It may be easy to assume replication and geographic separation are part of a service, but doing so may put your collection at risk.

Questions to ask:

- Is data backed up and redundant across at least two geographically separate data centers? Where are they located?
- What are the storage types used by the service (tape, disk, combination)? What does this mean for hardware updates and data migration?
- What is the service provider's hardware update protocol? How often do they update hardware (replacing drives and/or tapes)?
- What checks and balances are employed when migrating data? Are results of hardware migration reported or available to the customer?
- Are hardware/media health reports relative to the location of the customer's stored data made available to the customer?
- Do they provide their service via an existing storage provider, such as Amazon, or do they own and maintain their own servers?

[1] For more information and recommendations on planning for digital preservation, consult the NDSA's Levels of Preservation. http://www.digitalpreservation.gov/ndsa/activities/levels.html

4. Data Management

In turning over some elements of data preservation to a third-party, there may be some trepidation about exactly what happens to your files once they enter a new environment that is out of your control. The vetting and exploration process should remove the opacity of a service provider's operation, where possible. It should relieve any unease that your collection is being put away in a black box and handled in unknown and unforeseen ways. A third-party service must, at minimum, be able to deliver back to you exactly what you put in.

Again, adhering to basic digital preservation principles is key. Vendors may be able to confirm the integrity of your files—verification of the current file state against a known baseline, such as checksums—upon delivery and through ongoing management of them while in storage. If so, how they do this is important. If not, it should be determined how management of file fixity and integrity may be handled by the client, and what this means for your operations and resources.

Questions to ask:

- Are any alterations made to the files upon ingest into the storage environment (e.g., encryption or de-duplication)?
- Does the service ensure the integrity of your data? If yes, how do they do it and how often?
- What happens if a data integrity check fails?
- How does the vendor track and monitor your data within their system? Do they generate any metadata that could be used in client-side operations? Do they offer reporting to the customer?
- What standards and/or certifications does the service comply with (e.g., ISO 16363 – Audit and certification of trustworthy digital repositories)?

5. Reporting & Metadata

The importance of documentation in digital preservation cannot be understated, but the reporting and metadata requirements of an archival collection may be foreign to most third-party storage vendors. Archives must be able to account for events occurring in a digital asset's lifecycle, track any changes to the data (desired or not) to ensure the integrity of collection materials, and troubleshoot any issues that may arise. Management of these factors is mitigated by development of preservation policies and procedures enacted by multiple parties, who are often not the policy maker. System reports can close the loop, providing the policy maker and collection manager with evidence of the proper enactment and efficacy of existing policies. This approach supports informed decision-making for future management of a collection and associated policies. The reporting and metadata features of the service you choose will have an impact on the resources required to provide ongoing access and preservation, and should not be overlooked.

Questions to ask:

- What reports and logs are available to the client (e.g., error logs, fixity check results, ingest results, migration reports, or access logs)?
- What metadata is generated during storage (e.g., fixity checks, ingest details, migration reports, or replication details)?
- What mechanisms (e.g. emailed reports, Admin panel or dashboard, API) enable a client to retrieve reports and/or metadata?

6. Performance

Having defined your information architecture, it must be determined whether a service can meet your performance benchmarks for delivery to and from storage. Each organization will have different needs, but the general terms of performance requirements will be determined by the purpose of the service you are considering. If being used for access, your requirements will likely necessitate high-speed delivery of materials from storage and the ability to support a number of users with a relatively low latency period. Conversely, preservation storage may not necessitate high-availability of materials, but in cases where access is required such as checksum validation or the creation of new derivatives, it is essential to know if the service places any restriction on access to materials in deep/lower cost storage. This could indicate whether envisioned access scenarios will result in additional charges for your organization (e.g., Amazon Glacier's additional charges for access exceeding **10TB** per month, or varying prices for varying retrieval periods). Some elements of these considerations will be influenced by technological factors at your organization (e.g., system architecture, connection and processing speeds, or number of potential users). All these factors must be taken into account to provide a full picture of your existing infrastructure abilities and needs.

Questions to ask:

- What is the maximum bandwidth permitted for delivery to/from the service?
- What is the estimated number of end users requesting access? Number of management staff accessing the service? What is the maximum number of concurrent users or connections accessing the system?
- What types of files will each user type be accessing (e.g. still images, documents, audio, video), and what types of access do they need (e.g. streaming vs. download)?
- Does the service provide delivery of assets to end-users, and if yes, how so (e.g., streaming, embed codes, or integration with systems via API)?
- What are your acceptable latency periods for retrieval and access, and what are the associated charges?
- How much data are you permitted to retrieve each month at no additional charge? What is the sliding pay scale once this limit is exceeded?

7. Security

Establishing trust in a vendor starts by assuring the security of materials handed over. This includes ensuring that sensitive information is handled appropriately and is safe from unwanted human intervention. It is up to the client to define their security requirements and the level of resources they are willing to apply to the ongoing management of security protocols. For instance, some services may provide security features like encryption, but it is important to know who is responsible for managing keys for decryption. In addition to a security consideration, this requires understanding the level of resources available for managing and ensuring the security of materials on the client side. When considering security options, it is important to also consider the impact of various security mechanisms on performance and permissions structures.

Questions to ask:

- Do the materials in your collection contain sensitive information that requires increased security in a cloud storage environment?
- What security protocols does the vendor have in place? Who is responsible for managing and maintaining keys?
- What security standards do they comply with and what certifications have they received (e.g., ISO 27001 – Information security management, HIPAA security standards for medical records, or Federal Risk and Authorization Management Program [FedRAMP])?
- Can the client apply and manage its own security protocols (e.g., key encryption)? What are the resource implications for managing in-house? Does the client have support for implementing and managing encryption and decryption keys?

8. Disaster Recovery

Closely related to security is the safety of your data in cases of disaster, whether man-made or an Act of God. There are plenty of horror stories of unexpected data center failures and it is important to consider the implications of these types of events, but it is even more valuable to focus on the vendor's response to disaster scenarios and assess notification methods, client service and support, and reconciliation plans. In most situations, the switch to backup should occur seamlessly. The client must understand how processes are executed to ensure materials are backed up and protected against loss, how the vendor has and will utilize methods of communication to report issues, and be assured that an infrastructure is in place to support the ongoing provision of the services as arranged. It is important for the client to make a distinction about protection of data and time-to-recovery. Each organization will want the ability to recover all data in the instance of a disaster and will have different requirements regarding the time it takes to recover their data following a catastrophic occurrence. These varying times for recovery can come with dramatically different price tags.

Questions to ask:

- What are the service's disaster recovery backup storage mechanisms (potentially in addition to a replicated copy)?
- What are the protocols for addressing loss of data and return of service in cases of disaster? What are the policies and details of the service agreement regarding loss of data?
- What are the service's communication methods and policies in cases of service outage and/or data loss?
- What is the maximum amount of time between when data is ingested into the primary storage location and when it is replicated to a geographically separate redundant storage location?
- What is the maximum time it takes from the point of disaster, or loss of the primary storage system to data to recovery and access to all data present prior to the disaster?

9. *End of Service*

Exploring the extent of a service includes a review of all potential scenarios in its lifecycle including termination by the client or the vendor. End of service agreements and protocols should specify the return of materials and metadata in full and as you submitted them; an organization should get back what it puts in (plus any metadata generated during storage). This process may be triggered by the termination of service from the vendor side as well, in cases where the client is unable to pay or the service will no longer be offered. Failure to identify these policies may risk loss of collection materials if the end of service, whether instigated by the client or vendor, is not understood. In addition to identifying the protocols, it is also important to understand the costs.

Questions to ask:

- What protocols and mechanisms are in place if the vendor no longer supports the service?
- How is data returned to/retrieved by the client upon termination of service?
- What costs are associated with removal of assets from the service?
- What actions lead to termination of service and how is this managed by the service provider?
- Does the service agreement clearly document this information in sufficient detail?

Conclusion

When considering a third-party service, keep in mind the criteria outlined above, but do not neglect the other factors required for implementation of a successful preservation environment. Cloud services will only take your organization so far.

A preservation repository is a combination of technological and human factors, automatic processes managed and implemented by manual input, and policy. The service you select will be the one that equips your staff with the tools to best support the preservation of your collection. Vetting is a time-consuming but valuable process. The resources put in will pay off with the successful integration of the right cloud service(s) in your preservation environment. To aid in the decision making process, AVPreserve will be releasing vendor profiles summarizing numerous services' abilities in relation to the nine important factors discussed above.

AVPreserve is a full service media archiving and data management consulting firm. We partner with Archives, Museums, Government Agencies, Corporations, Media & Entertainment, and other organizations that create or collect media to help them manage, access, and preserve their valuable assets and data. Our services address the full lifecycle of collections, from assessment and preservation planning for analog materials, through project management of digitization efforts, to the various aspects of digital preservation and file management, including DAM selection, taxonomy development, policy and workflows, and development of software solutions supporting preservation and access.