Journal of Archival Organization

How to Talk to IT about Digital Preservation

--Manuscript Draft--

Full Title:	How to Talk to IT about Digital Preservation
Manuscript Number:	WJAO-2018-0024
Article Type:	Column
Keywords:	digital preservation, professional communications
Order of Authors:	Scott Prater

How to Talk to IT about Digital Preservation

Abstract

When an archivist talks with the information technology personnel about providing access to digital collections, the discussion often proceeds smoothly. They have shared assumptions, the common goal of the platform is well understood by both parties, and the vocabulary used to describe actions, tasks, and items is substantially the same, or at least agreed-upon. However, when the discussion turns to digital preservation, there may be a divergence in priorities and understanding. While the archivist's priority is on making sure their digital assets are preserved and accessible *forever*, the IT personnel's focus may be on making current data publicly accessible, making sure that systems are running smoothly *right now.* To further complicate matters, IT personnel have been performing an activity that sounds very much like digital preservation: archiving, backups and storage of multiple copies. This column will explore the convergence and divergence of the archival and information technology professions in regards to digital preservation, as well as suggest the topics and questions that need to be discussed so that both parties have a mutual understanding of meeting the needs of providing appropriate digital preservation.

Introduction

One of the most difficult conversations an archivist can have with their organization's information technology (IT) staff regards digital preservation. The archivist may present IT personnel with what seems a relatively straightforward request for storage to preserve files. IT may respond with a request for more information, often couched in language that is both familiar and unintelligible. The archivist becomes frustrated because IT seems to be stonewalling them,

Submission for Journal of Archival Organization, Technology Matters in Archives Column, Volume 15, Issue 3

Title: How to Talk to IT about Digital Preservation
Author: Scott Prater
Digital Library Analyst
UW Digital Collections Center
University of Wisconsin - Madison

and the IT staff throws up their hands because they cannot get the answers to what seem to them perfectly obvious questions. The conversations frequently go south from there, if they get started at all.

More than likely, the archivist already has a workflow and infrastructure in place for the search-and-discovery platform, the public website where carefully curated digital collections can be found and referenced by the general public. This site is up-and-running, has been for several years, has solid IT infrastructure support behind it, and is scaled to handle its primary job: making materials available to the public. For IT staff, this is a known, recognizable environment; it is a website, a database, or another application they manage that is designed to handle current data in real time for active users. When the archivist talks with the IT personnel about the digital collections, the discussion proceeds smoothly. They have shared assumptions, the common goal of the platform is well understood by both parties, and the vocabulary used to describe actions, tasks, and items is substantially the same, or at least agreed-upon.

Preservation, however, is new territory, especially to rank-and-file information technologists who count the archives as only one of the many constituencies they serve. Most IT shops strive to standardize processes and infrastructure for all their users, to increase efficiency, reduce duplication of effort, and free up time and resources to improve their suite of services and introduce new ones. Their focus on is making current data publicly accessible, making sure that systems are running smoothly *right now*. Up to now, digital preservation services have not been on the list of standard services an IT shop has offered, and consequently, have received very little, if any, attention from mainstream system administrators (sys admins) and software developers.

To further complicate matters, sys admins have been performing an activity that sounds very much like digital preservation: archiving, backups and storage of multiple copies. This activity usually takes place in the background, invisible to end users. The sys admin takes pride in hiding backup and storage processes from view, so that ideally storage problems are never even noticed by users when they occur. They may also feel some pride of ownership in the domain, and may react poorly to what they perceive as outsiders meddling in their job. Mitigation of risk with backups and multiple copies is a keystone service of the sys admin's profession. Systems backup is a field with its own long, rich history, with an established body of knowledge and practices that sys admins justifiably consider one of the more notable mysteries of their profession. Yet there are aspects to strong digital preservation, which are not typical of your average backup system that archivists need to share with IT administrators.

Just as archivists have a language specific to digital preservation, IT has its own for backups and storage. These languages can sound awfully similar as the two share common roots. However, the use of key terms, such as the loaded word "archive," can differ in subtle and crucial ways that later cause confusion. Further, as would be expected, the IT language has a host of technical terms and acronyms that may be obvious to the cognoscenti, but meaningless to the non-technical professional, such as RAID, WORM, deduplication, SAN, and NAS among others. Although a detailed technical understanding of these terms is not necessary, an archivist who has some familiarity with this vocabulary will be in a much better position to make their needs understood.

Finally, the differing expectations between end users and IT support frequently come into conflict in regards to digital preservation and storage. Users are accustomed to having all the space they need, when they need it; in most cases, a simple request for more space is enough to get more space. Most users also assume that their data is protected and will always be available to them for the foreseeable future. The systems administration profession has worked hard to create, and then meet these expectations; and for the needs of most end users, the systems in place are enough. However, for digital preservation purposes, the bar is higher, in terms of both the quantity and the specific qualities of storage. What may seem like a simple request from an archivist, "I need 20 terabytes of preservation storage for this set of materials," may actually present a real problem for many IT shops. Worse, it may not; the IT shop may think it knows exactly what the archivist wants, that it is a simple request, and provide a storage solution that meets their understanding of digital preservation, but not the archivist's.

How can we talk with our IT support to develop a digital preservation infrastructure that they can understand and support and that meets our requirements? The rest of this column provides some suggestions for making the discussions useful and productive.

Context and Vocabulary

Here is what remote allocated disk storage space looks like to an end user:

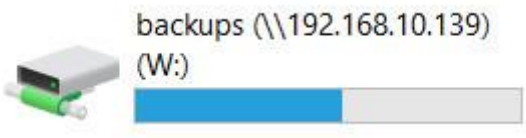

Figure 1: Remote allocation storage from user's point of view

Here is what it looks like to the sys admin who manages the W: drive:

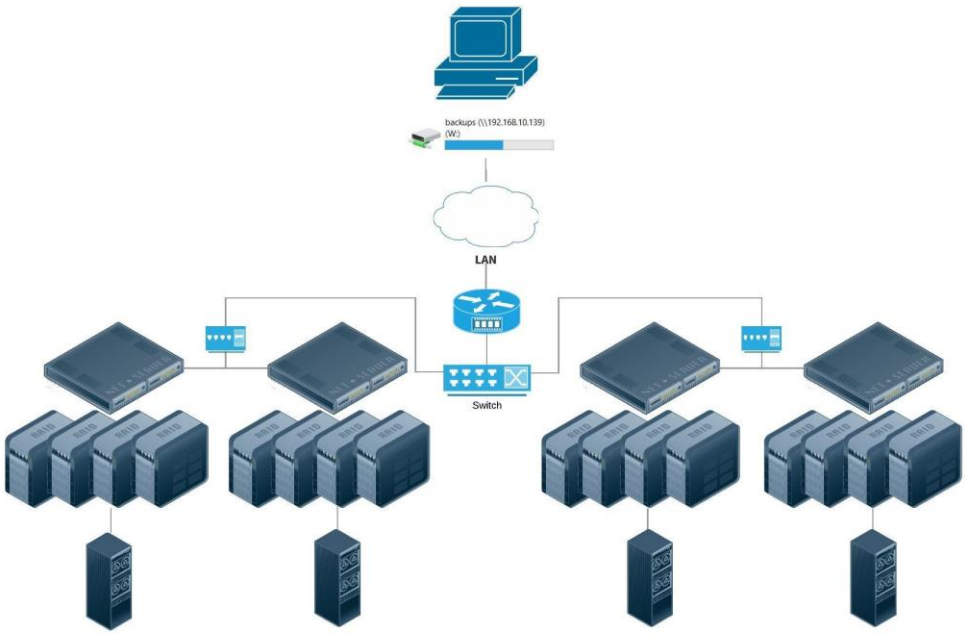

Figure 2: Remote allocation storage from sys admin's point of view

As this diagram shows, a system administrator's view of storage is considerably more nuanced than the average end user's. When discussing storage with IT staff, it is helpful to both present the request in terms that they can understand, and to have some knowledge of how they think of storage. Towards that end, below is a simple vocabulary list of the most common IT terms used when discussing storage, the meanings of the terms, and their relevance to an archivist concerned with digital preservation. The list begins with the most basic pieces of a storage system, then builds upon them to describe the components of a more sophisticated environment.

Archive: In the IT world, an archive is a set of files, usually accompanied by a manifest, that are bundled together into a single package (file). The manifest contains technical metadata about the files: names, folders, computer user name of the owner, permissions, size, when the file was last modified, etc. Archives are a point-in-time snapshot of the set of files, usually designed to be stored offline on some medium (e.g. tape, CD, or cloud storage provider), so that the archive later can be unpacked, and the files and folders restored to the state they had at the moment the archive was made. A zip file is a kind of archive.

> *What this means to the archivist*: An archivist's archives and IT archives are not the same thing, although they may have considerable overlap. How this word is to be used in discussion should be explicitly made clear up front.

RAID: Redundant Array of Independent Disks. This is a set of disk drives linked together to form one big virtual disk drive in a computer. They are usually grouped together in such a way that data is written across all the disk drives, and that the failure of any single disk and subsequent loss of data can be recovered by restoring data written on the other disks. There are many ways a RAID can be organized, and they are usually referred to by a number extension such as RAID 1, RAID 2, RAID 3, etc.

> *What this means to the archivist*: Truthfully, not much. RAID is a way to mitigate the risk of data loss when a physical drive fails. IT administrators take care of this, and rarely, if ever, will the archivist need to be concerned with the RAID organization of the disks storing the data. However, it is a very common IT storage term, and an archivist's main

concern should be to understand how the array is configured, and how its setup minimizes the risk of catastrophic loss.

NAS: Network-attached Storage. This is a computer with many disks, linked together and organized in a RAID setup, then connected to a network. Its primary purpose is to group together all the disks in the computer into a seamless whole, then present that to users (or a SAN) over the network as one big remote drive.

What this means to the archivist: The physical bits may wind up living on a NAS; it may be a NAS where storage is allocated and grown for the archives. Understanding where the NAS physically resides, how it is connected to the network, how it is backed up, how it is made redundant, will help answer core digital preservation audit questions about the risks the data are susceptible to, and their mitigation.

SAN: Storage Array Network. A network of storage devices (e.g. tape drives or NAS devices) linked together in a larger network and presented to the end user as a set of remote drives. The diagram above shows a simple SAN. Most enterprise storage is managed in one or more SANs; system administrators often refer to the SAN as the entity where storage actions take place. They may say, “We can allocate 20TB for the archives on the SAN, then mount it as a shared W: drive.”

What this means to the archivist: The SAN is where most storage policies regarding multiple copies, disaster recovery, point-in-time snapshots, and file corruption are implemented at the physical hardware level. Multiple copies of files can be stored and

linked together in a SAN. Backups may be made in a SAN. A local SAN may often mask external cloud storage. Understanding at a high level the architecture of the institution's SAN allows the archivist to answer such digital preservation questions as, "Are there multiple copies of the assets? If so, how many? Where do they geographically reside? Where are the backups of the data made?"

Redundant storage: Storage arranged in such a way that multiple copies of data are created and updated in real-time or close to real time. This usually manifests itself as two or more SANs linked by a network connection, so that data written on one device in SAN 1 is automatically written to a matching device in SAN 2. The individual SANs in a redundant storage setup are called *nodes*; the act of copying data from one node to another is *replication* or *mirroring*.

What this means to the archivist: How the storage is set up for redundancy answers the question of how multiple copies of the data, with geographic distribution are configured. Additionally, there are a couple of things for archivists to keep in mind regarding redundant storage:

- It is NOT the same as an offline backup. A corrupted file written to SAN 1 will also be written to SAN 2; a file deleted on SAN 1 will also be deleted on SAN 2. Redundant storage can be used to mitigate the risks of bit rot (if bits flip on a file in SAN 1, the copy on SAN 2 can be used to fix the SAN 1 file), and to mitigate the risk of hardware failure in SAN 1 or SAN 2. It *does not* mitigate the risk of errors or corruption introduced when the archival staff is working at their workstation(s), writing and copying and deleting files stored in the remote SAN.

- Multiple copies are just that: copies. This means that the 10 TB of data, in a two-node SAN with redundant storage, actually takes up 20 TB of disk space. Add more nodes, and the amount of space rises accordingly. Then throw into the mix an offline tape backup, with point-in-time snapshots archived off for a year, and the 10TB can balloon into 50 TB or more of disk and tape space consumed.

Checksums: A checksum is an alphanumeric string generated by an algorithm on a file or set of files that acts as a unique identifier or "fingerprint". The value is used to determine and verify the *fixity* of a file or set of files. If a file's checksum changes, it indicates that something has changed in the file, and may be an indicator of file corruption.

What this means to the archivist: *Fixity* is not a common term in systems administration; however, by explaining it in terms of checksums, IT personnel will understand the meaning. Some of the more advanced SAN systems manage regular fixity checking as a method of "self-healing." If the SAN has a self-healing system, the archivist should ask if they can get reports of when fixity checking happens, and when corruption is detected and addressed.

WORM: Write Once, Read Many. This is a type of storage that enforces versioning, by making it impossible to overwrite a file. Any time a change is made, a new version of the file is created on disk. Most cloud storage providers offer some kind of WORM storage.

What this means to the archivist: Systems that offer WORM storage enable versioning at a low level (bits on disk) for data to be preserved. However, this alone may not be

enough for full digital preservation versioning, as generally no metadata about the change event can be stored at this level.

High-availability: High-availability storage is optimized for access; it is fast, and has a high degree of redundancy to make sure it never goes down. It is also more expensive.

What this means to the archivist: Space for public access digital collections should have high availability, but it is not necessarily required for the offline storage of large preservation-quality objects. Opting for a low-availability preservation storage system that need not be publicly accessible in real time provides a more economical strategy for digital preservation.

Deduplication: Most modern SAN systems can detect when bytes are duplicated on its disks, and store just one copy of the bytes on a node, with multiple pointers to that copy. Deduplication is the internal process of finding duplicate bytes and replacing multiple copies with just one. Note that this type of deduplication works at the system level, and should not be confused with removing redundant copies of files.

What this means to the archivist: Depending on the nature of the stored data, this can reduce the amount of physical disk space needed, and hence, storage costs. A system that supports deduplication can reduce 20TB to 15TB, for example, if many of the files are identical, or even mostly identical, such as the white space in a document.

Backup: System administrators understand backups as copies of data made periodically then stored offline, and preferably offsite. The purpose of a backup is to restore the data on a system to a point in time before the data became corrupted or disappeared. The copies of data are kept for a period of time, then should be erased or overwritten.

> *What this means to the archivist*: Backing up files is not the same as preserving them. Backups are solely designed to mitigate the risk of data loss in the event of an accident or attack, providing for business continuity. They are not designed to store data permanently offline for later retrieval. However, many of the tools and techniques used for making and maintaining backups can also be repurposed for digital preservation activities.

Keep in mind that sys admins responsible for storage think in terms of bytes, blocks (fixed groups of bytes), files, and directories. They are responsible for storing those things on disks, which are grouped together into servers that are linked by a network. The more an archivist can translate digital preservation concepts and entities into concrete actions performed on files on a disk, the easier it will be to make archival needs understandable to IT support.

The Conversation

Armed with a general idea of how storage systems work under the hood, an archivist should be in a good position to begin talking about digital preservation storage in terms that are tailored to the mindset of IT support. Before initiating a discussion, however, an archivist should be prepared to answer the following questions.

- What is the space going to be used for?
    - Distinguish between the amount of space needed for ongoing work, the space needed for making the source materials or derivatives publicly available, and the space needed for permanent storage.
    - If the plan is to separate the preservation objects from the objects in the publicly-accessible digital collections, the archivist needs to be prepared to explain briefly why that approach is being taken. It is possible that IT support may consider the digital collection platform the solution to the preservation problem, and may need further explanation as to why that is not the case.
- How much data will be processed and stored?
    - It is best to have a general idea of both the number of objects, and the size of the files. Some storage systems are optimized for a few large files, others for many small files.
    - Further, the archivist should be prepared to discuss storage growth projections.
- At what rate will the data be accessioned or reformatted, ingested, and preserved?
    - Knowing this will enable the archivist and IT support to plan, and prioritize the storage setup tasks.
- Does the preservation storage need to be high-availability, with the data immediately accessible, or can it be retrieved within an acceptable time period?
    - If it does not have to be high-availability storage, what is an acceptable amount of time to wait for retrieval?
- What type of controls can be implemented to manage access privileges to the data?
- What are the requirements for:

    - multiple copies
    - versioning
    - offline point-in-time snapshots
    - geographic distribution of copies
    - fixity
    - preservation activity audit trails
- What is the processing workflow?
    - How does digital data arrive?
    - Where is it processed?
    - How does it get into the presentation platform?
    - What files belong in the preservation platform?
    - At what point in the workflow are files committed to the preservation platform and can be removed from the processing area?

When archivists have a good, detailed understanding of their own needs, a request can be framed that IT support can respond to. Follow-up questions to consider asking the IT support personnel:

- How, concretely, will the requirements be addressed?
    - Explain to IT that for auditing purposes, there is a need to document the digital preservation strategy in great detail. They can help concretely answer the questions posed by an audit.
- What are the storage options, based on the needs articulated by the archivist?
    - What are the pros and cons for the options?

- What are the costs associated with the various storage options?
    - The total cost of ownership of on premise storage includes not only the hardware and software, but also overhead of plant operations and maintenance, as well as personnel costs.
    - Cloud storage vendors can provide a multitude of services. Cloud storage is becoming increasingly popular, and for good reason: its costs are coming down, its service offerings are improving and expanding, and it takes a huge maintenance burden off IT support. The archivist and IT personnel need to determine if cloud storage services are the most cost effective, and which requirements of preservation storage they meet. See below for more cloud storage considerations.
- What is the backup strategy?
    - Are backups stored offsite?
    - Is the offsite storage geographically dispersed?
    - Are multiple copies of the backup set made?
    - How often are backups performed, and how many backup sets are kept, for how long?
    - How far back in time can you go to recover data?
    - Has a successful restore test been actually completed?
    - Keep in mind that backup strategies are for business continuity, not preservation. As such, they have a rolling time window of data that can be restored, and that earlier versions of data should disappear after a certain period.

- When considering cloud storage as a piece of a larger preservation strategy, not as a one-stop shopping solution for all the institution's preservation storage needs, ask these questions:
    - How is fixity managed, if at all?
        - If the cloud storage service offers fixity checks, does the institution have access to the checks and results?
    - How does the cloud provider manage versioning?
        - What implications does this have for storage costs over time?
    - How easy will it be to perform format migrations, or other large-scale transformations of data stored in the cloud?
    - What are the network upload costs?
    - What are the metadata viewing costs?
    - What are the costs for downloading data stored in the cloud?
    - What is the institution's exit strategy, if the service goes down, the provider closes shop, or the terms and costs of service become unacceptable?
    - Does the vendor provide information as to where the data is physically stored?
        - Some types of data have regulatory constraints about overseas distribution.

Finally, be ready to negotiate. An archivist needs to consider what is necessary to preserve into perpetuity, and what has a limited lifecycle and should be disposed of at the appropriate time. Ideally, we would always save all source materials and derivatives for all time, but realistically, there are not enough fiscal and human resources to accomplish this, nor is there necessarily a demand to do so.

Probably the most important question an archivist needs to ask is, “When is it time to begin the conversation with IT support?” The answer is, “As early as possible.” As soon as the archivist has enough concrete information about an upcoming project or set of materials that need to be stored and preserved, it is the time to begin involving IT in the planning. Although the exact details may not be readily available, communicating the general contours of the upcoming project and engaging IT support early on can lead to more clearly defined requirements, and will go a long ways towards ensuring a smooth, successful project. For example, when writing a grant, have this conversation while developing the proposal, so that concrete expectations, deliverables, and costs can be worked into the grant.

Most of the questions and concerns outlined above are covered in *The National Digital Stewardship Alliance (NDSA) Levels of Digital Preservation*[1]. This document contains a useful chart to share with your system administrators, and can provide a framework for discussing needs and expectations, and the rationales behind them.

Conclusion

How not to ask for storage:

[1] Megan Phillips, Jefferson Bailey, Andrea Goethals, and Trevor Owens, “The NDSA Levels of Digital Preservation: An Explanation and Uses,” *Proceedings of the Archiving (IS&T) Conference*, April 2013, Washington, DC. http://www.digitalpreservation.gov/documents/NDSA_Levels_Archiving_2013.pdf (accessed March 11, 2018).

Help! I need 5TB more of disk space on the S: drive ASAP!

How to ask for storage:

I'd like to sit down with someone managing storage services to talk about allocating disk space for an upcoming project that we are contemplating for the fall. During the course of the project, we would receive on external hard drives approximately 20TB total of uncompressed audio files over the course of six months. We would need an area where we can move these files off of the external hard drives, then process them for our digital collections and preservation storage. The estimated space we would need for processing is 5TB. We would need another 5TB for our generated digital collection files, and then about 20TB of preservation storage for our source uncompressed master files. Let's arrange a time to meet and discuss this request, and identify what our requirements and options are.

Knowing what the needs are, and being able to explain them in terms that IT staff are familiar with, is 90% of a successful conversation.

What is true in life is also true in digital preservation discussions: generosity, mutual respect, and patience win the day. An archivist should frame requests as interesting problems they would like to work with IT to solve, not as demands to be met. Think of IT personnel as collaborators, not just as service providers. Remember that digital preservation may be a new field for them, one that they may think is familiar, but that has some important nuances that it is the archivist's task to help them understand. Conversely, be aware that storage is not a

commodity that magically appears when needed, but is a complicated system that requires an entire profession to manage responsibly. Listen when IT personnel explain why a request may present challenges. If something is not understood, on either side of the conversation, stop the conversation and clarify.

Approached in this way, the initial conversation leads to more conversations that are characterized by mutual intelligibility, education and support. Fruitful discussions are the foundation of a lasting partnership with long-term benefits to both the archivist and the IT staff.

Figure

Figure