Fundamentals of AV Preservation

CHAPTER 4

MANAGING DIGITAL AUDIOVISUAL COLLECTIONS

Northeast Document Conservation Center | NEDCC
Andover, Massachusetts | 2022

PREFACE

NEDCC is pleased to offer the Fundamentals of AV Preservation textbook for self-study to anyone with internet access. An HTML version is available online at *www.nedcc.org/av-textbook*, and downloadable PDFs can be found at *www.nedcc.org/publications*. Covering the core topics in caring for and reformatting audiovisual collections, this resource supports cultural heritage professionals in their efforts to steward audiovisual materials.

- Chapter 1: Care and Handling of Audiovisual Collections
- Chapter 2: Inventory and Assessment
- Chapter 3: Planning, Preparing, and Implementing Reformatting Projects
- Chapter 4: Managing Digital Audiovisual Collections
- Chapter 5: Disaster Preparedness and Response
- Glossary

Credits

The content for each chapter of the Fundamentals of AV Preservation textbook was created in 2017 by staff members of AVPreserve (now AVP) and edited by NEDCC staff members. The textbook served as the foundation of a multi-session, instructor-facilitated online course launched by NEDCC in 2017. The development of the course and textbook were subsidized by the National Endowment for the Humanities, and in 2022, additional NEH funding supported the transformation of the HTML textbook into downloadable PDFs.

AVPreserve — Authors
Chris Lacinak, *President (Ch. 1 & 3)*
Rebecca Chandler, *Consultant (Ch. 1 & 2)*
Amy Rudersdorf, *Senior Consultant (Ch. 4)*
Kara Van Mallsen, *Partner and Senior Consultant (Ch. 5)*

NEDCC — Editors
Ann Marie Willer, *Director of Preservation Services*
Sean Ferguson, *Preservation Specialist*
Becky Geller, *Preservation Specialist*
Frances Harrell, *Senior Preservation Specialist*
Kim O'Leary, *Technology and Events Coordinator*
Danielle Spalenka, *Preservation Specialist*

CONTENTS

CHAPTER 4

MANAGING DIGITAL AUDIOVISUAL COLLECTIONS

by Amy Rudersdorf, Senior Consultant, AVPreserve

THE FUNDAMENTAL REQUIREMENTS OF DIGITAL preservation are threefold: (1) maintain the bits, or building blocks, of the digital files; (2) maintain the content of the file (the movie in the video file, the song in the audio file) so that it is accessible and understandable; and (3) preserve both the bits and content for as long as necessary. This last requirement demands more than a technological solution. Even more than physical audiovisual collections, digital collections require holistic management to ensure their long-term preservation and access—and organizational factors are critical. Incorporating the organization's goals and objectives is essential to ensuring support for a sustainable digital preservation program. Without high-level organizational buy-in, ongoing resources (staffing, funding) together with technology (hardware, software, storage)—essential aspects of a sustainable digital preservation program—are not guaranteed. Digital audiovisual collections management involves planning in order to establish policies and standards-based practice, so that staff understand their roles and so that technology can be best utilized for an institution's digital collections.

The world in which digital collections live is one of constant change. For audiovisual files, carriers (storage media) and wrappers and codecs (together, the file format) must each be addressed separately when thinking about best practices for long-term management. Change can occur at any level. For example, a digital storage system may need to be replaced even if the file format within it is stable. Likewise, a file format may need to be migrated to a newer version, while storage remains unchanged. In the digital world, where every bit and byte matters, keeping in mind the relationships between file, carrier, and storage will help you best manage your collections.

This chapter will include an overview of the approaches, standards, and considerations for an institution beginning to manage their digital collections. The beginning of the chapter focuses on the organizational infrastructure that supports sustainability of digital collections. Organizations must consider factors that make long-term preservation possible, such as risk management, preservation management, standards and guidelines, policies and planning, organizational infrastructure, and planning and phasing. The second half of the chapter describes the activities and technology required for digital preservation. These include storage (redundancy, media, and geographic diversity), monitoring (fixity), information security, metadata, refreshing and migration.

When establishing a digital preservation program, keep in mind that not all of the activities in this chapter need to be put in place at once. For some institutions, implementation of a complete set of digital preservation policies, for example, may not occur right away. Building momentum within the institution might be the first step, which will ultimately lead to the adoption of policies, committed financial support, and deployment of technologies. The recommendations herein should be considered the

fundamental components of a sustainable preservation program. With continued evolution and improvement, the adoption of these plans, policies, and services will support the goal of long-term preservation and access to your digital audiovisual collections.

SECTION 1: DIGITAL PRESERVATION CONCEPTS

Definitions

"Digital preservation is the active management of digital content over time to ensure ongoing access."[26] It is an integral part of a larger process of **curation**, which consists of the activities across the content lifecycle described throughout this textbook: selection and appraisal, description, ongoing care and management, long-term access, and/or deaccessioning/disposal.

Without **active management**, which includes many of the activities outlined in this chapter, digital assets and their associated content are at risk, potentially through media failure, human error, inaccessibility due to format obsolescence, or barriers to discoverability due to poor metadata. The functions of digital preservation reduce these risks and together help to ensure that content remains accessible over time.

Risk Management

At the heart of a preservation strategy is **risk management**. The opportunities for loss or damage to digital collections is inherent in providing access to users, who may include digital collections managers, IT staff, students, the public, and others. A preservation strategy can be an effective way to mitigate these risks to the greatest extent possible.

The nature of risks is varied and may be human-generated, mechanical, or natural. The human risks to technology may be purposeful (file formats are not selected for **migration**, metadata is not captured), nefarious (viruses, cyber-attacks), or accidental (deletions, misfiling or misnaming files). Organizational risks include insufficient planning and policies, which lead to a loss of or lack of sustainable funding to support trained staff and/or appropriate technologies. Risks may also be mechanical, such as when files change at the bit level without human awareness or media and storage fail. Risks may also come from nature; floods and fires can destroy electronic media on which files are stored.

Over time, risks evolve based on the organization and its resources as well as industry-wide technical changes. As risks change, how institutions identify, respond, and monitor them must change, too. Successful preservation strategies must be flexible, yet cautious, to be able to react to risk effectively. Through planning and management, risks can be mitigated as they emerge.

Preservation Management

Due to the varied nature of risks to digital assets, it is important that digital preservation be approached from a programmatic standpoint with administrative support that makes preservation a priority. One approach to thinking about preservation management is illustrated by the concept of the "three-legged stool," in which the organizational infrastructure, technological infrastructure, and resources all have equal footing to create a stable digital preservation program. Nancy McGovern's *Digital Preservation Management: Implementing Short-term Strategies for Long-term Problems* describes these three interlocking structures as follows:

1. *Organizational Infrastructure* includes the policies, procedures, practices, and people—the elements that any programmatic area needs to thrive but that are specialized to address digital preservation requirements.

2. *Technological Infrastructure* consists of the requisite equipment, software, hardware, secure environment, and skills to establish and maintain the digital preservation program. It anticipates and responds wisely to changing technology.

3. A sustainable *Resources Framework* addresses the requisite startup, ongoing, and contingency funding to enable and sustain the digital preservation program.[27]

26 Library of Congress, "What is Digital Preservation?" *http://www.digitalpreservation.gov/about*

27 Anne R. Kenney, et al. "Digital Preservation Management: Implementing Short-term Strategies for Long-term Problems." *https://dpworkshop.org/dpm-eng/conclusion.html*

Determining the "what" (organizational infrastructure), the "how" (technological infrastructure), and the "how much" (resources framework) helps to identify an institution's needs for a sustainable digital preservation program.

Standards

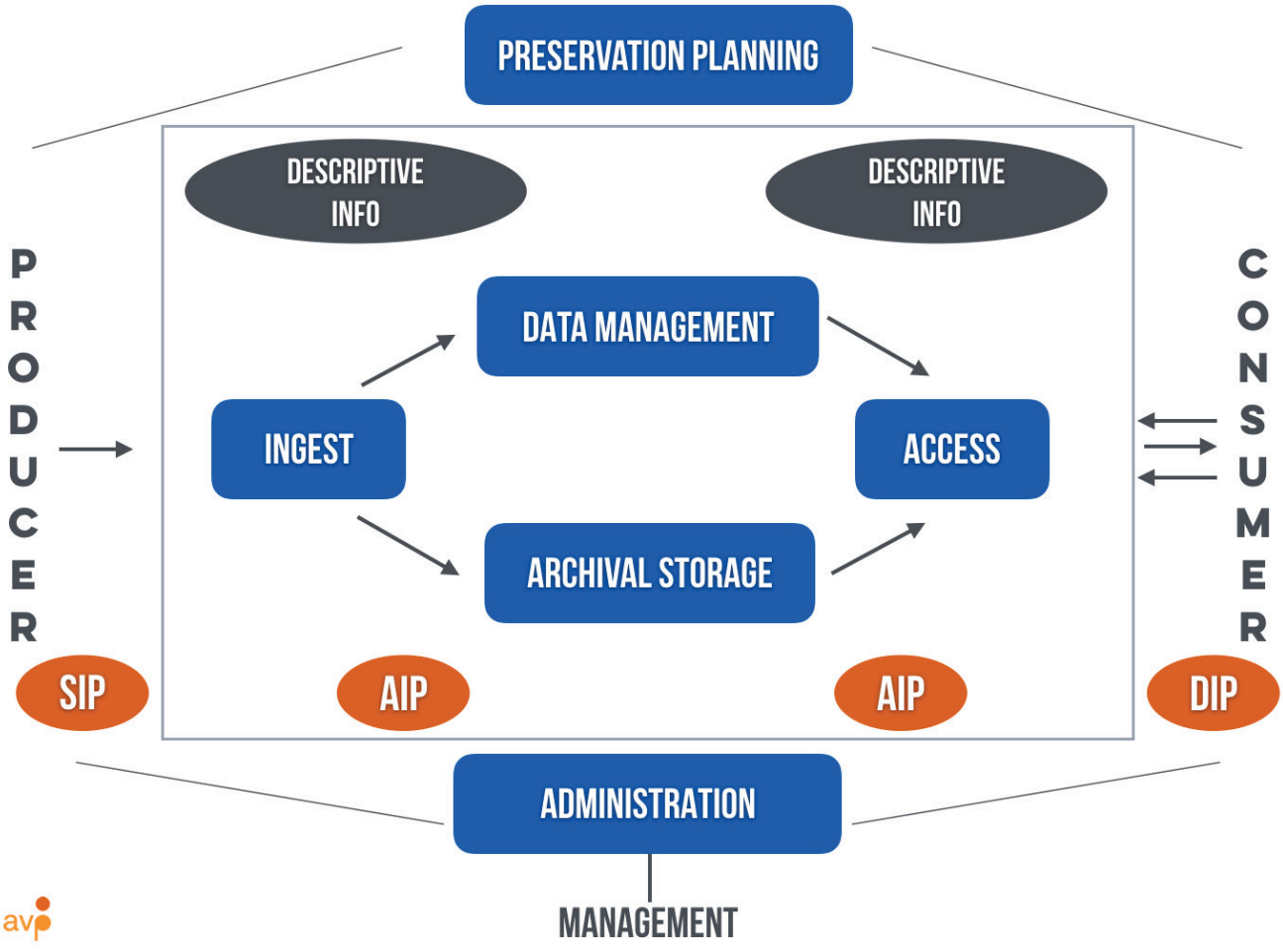
Many industries employ standards to make certain they comply with accepted practice, ensure the safety of their customers and employees, and provide a foundation upon which new technologies can be built. Two international standards documents serve as the cornerstone for the management of digital collections, guiding institutions in the development of sustainable preservation programs and serving as benchmarks for institutions that are maintaining preservation management technologies. These are:

- *ISO 14721:2012, Space data and information transfer systems—Open archival information system (OAIS)*—Reference model, commonly referred to as the OAIS Reference Model, or OAIS, and;
- *ISO 16363:2012, Space data and information transfer systems—Audit and certification of trustworthy digital repositories.*

OAIS REFERENCE MODEL

The Open Archival Information System (OAIS) reference model is a conceptual framework for an archival system dedicated to preserving and maintaining access to digital assets over the long-term. Because it is a conceptual model, it is not inherently prescriptive. Instead of specifying technology, staffing, or resources, it provides guidance

FIGURE 4.1
Open Archival Information System (OAIS) Reference Model
Credit: AVP



about best practice for building a sustainable preservation environment. OAIS states that archival storage is necessary but does not specify how that storage will be implemented, the technology required, or how many staff are needed to maintain it. Likewise, access is a component of the OAIS model, although how access is provided will be unique to each organization. Each component represents a process within a system but not the specific resources and technology required to support that process.

The framework takes into account producers (creators) of content (and their embodiment as file-based assets and data that will be preserved); the system (technology, workflows) in which the content is preserved; the administration, management, and preservation planning structure that administers the program; and the consumers (users) that will use the content at some point in the future. Content is packaged in different formats throughout its lifecycle (reflected in the orange ovals in the graphic on page 61):

- Submission Information Packages (SIPs), created in preparation for **ingest** (submission into the archival system)
- Archival Information Packages (AIPs), the content that is managed and stored over time, which may include the SIP contents and additional data created by the system
- Dissemination Information Packages (DIPs), the content made sharable for users, typically a subset of the AIP

The value of OAIS is that it provides a model for functions that should occur in a preservation environment and the types of content that must be managed over time. It is an example of a holistic approach to digital preservation, taking into account not only the technology but also the people, resources, and organization as well. Compare this conceptualization to the three-legged stool analogy above.

AUDIT AND CERTIFICATION OF TRUSTWORTHY DIGITAL REPOSITORIES

ISO 16363 is the international standard that describes the characteristics of a trustworthy digital repository (TDR). It includes categories of metrics that identify the individual components that together comprise a TDR. The OAIS framework is the basis for the TDR. The expectation is that any TDR will embody the OAIS model, including a robust submission and ingest process, an archival storage and data management system, and an access component—all overseen by a fully developed organizational and management structure that ensures programmatic longevity and stability. See the figure above to see how these facets interact. Digital objects in ISO 16363 are referred to as "information packages," as in the OAIS standard. Other vocabulary from OAIS appear in the standard as well.

While the standard was developed to provide a framework for certifying a digital preservation program as "trustworthy," the reality is that the great majority of institutions will probably not attain certification. Instead, many organizations use the standard's 109 metrics to guide development and growth of their program and archive and to focus energy and resources on areas for improvement.

STANDARDS-BASED DECISION MAKING

Standards offer comprehensive guidance on the making of highly functional and robust digital preservation environments. However, implementing these standards can be a major undertaking, requiring significant resources and cooperation from numerous stakeholders. If an organization lacks the resources to build a local system that meets these standards, there are a number of alternatives available to it, but these standards should still be consulted to provide guidance in choosing the right option.

Look for partnership organizations. Many, often larger, organizations have built their digital preservation infrastructure with standards in mind. Partnering with them to deposit into their archive might be a more sustainable option than building your own. Review a partner organization's documentation and talk to its staff to understand their approach to digital preservation and whether it complies with OAIS and TDR.

Consider consortial or partnership organizations. Organizations like DPN, MetaArchive, and others have built communities around their digital preservation services. They work together to ensure that these services are built within the ISO framework.

Purchase preservation storage services. DuraCloud and Preservica are just two of several digital preservation products built with guidance from the ISO standards.

No matter which direction an organization takes—building its own preservation system or looking outward for services —the ISO standards should be kept in mind. Whatever the approach, standards-based decision making will help to build a robust and sustainable digital preservation program.

Resources

Definitions

- Association for Library Collections and Technical Services. "Definition of Digital Preservation." *http://www.ala.org/alcts/resources/preserv/defdigpres0408*
- Digital Curation Centre. "What is Digital Curation?" *https://dcc.ac.uk/about/digital-curation*
- *Digital Preservation Handbook.* "Digital Preservation." *https://www.dpconline.org/handbook/glossary#D*
- Library of Congress. "About Digital Preservation." *http://www.digitalpreservation.gov/about*

Standards

- ISO 14721:2012, Space data and information transfer systems—Open archival information system (OAIS)—Reference model. *http://public.ccsds.org/publications/archive/650x0m2.pdf*
- ISO 16363:2012, Space data and information transfer systems—Audit and certification of trustworthy digital repositories. *https://public.ccsds.org/pubs/652x0m1.pdf*

Guidelines

- Data Seal of Approval. *https://www.coretrustseal.org/about/history/data-seal-of-approval-synopsis-2008-2018*
- "Levels of Digital Preservation." National Digital Stewardship Alliance (NDSA). *http://ndsa.org/activities/levels-of-digital-preservation*

SECTION 2: ORGANIZATIONAL INFRASTRUCTURE

A key facet of a sustainable preservation program is the organizational infrastructure that supports it. This includes all the elements that any programmatic undertaking needs to thrive, including planning, policies, funding, procedures, stakeholders, and decision makers, all of which must be tailored to address digital preservation requirements.[28] Organizational infrastructure is one of the three main components of the international standard described above, *ISO 16363: Audit and Certification of Trustworthy Digital Repositories,* which defines programmatic **governance** structures (i.e. dedicated leadership and oversight functions), preservation policies, and sustained support for staffing and funding as fundamental.

To establish organizational support for a digital preservation program, planning and documentation are essential. Documenting the current state of the collections and the future goals for preservation will help cultivate organizational buy-in for a digital preservation program.

Preservation Planning

The first step in developing a sustainable preservation program, or in other words, achieving the balance of the three-legged stool, is planning. Preservation planning can be defined as "a process by which the general and specific needs for the care of collections are determined, priorities are established, and resources for implementation are identified."[29] Digital preservation plans differ amongst institutions, but generally they support the following objectives:

- Organizational commitment to the preservation of and continued access to digital collections through ongoing financial and resource support
- Authenticity of digital collections (ensuring files are trustworthy representations of their original content)
- Environmental controls on the physical media on which digital collections are held

28 Anne R. Kenney, et al. "Digital Preservation Management: Implementing Short-term Strategies for Long-term Problems." *https://dpworkshop.org/dpm-eng/conclusion.html*

29 NEDCC, "Preservation Leaflet 1.1: What is Preservation Planning?" *https://www.nedcc.org/free-resources/preservation-leaflets/1.-planning-and-prioritizing/1.1-what-is-preservation-planning*

- Monitoring collections and addressing errors or changes as necessary
- Migrating digital collections as software and/or hardware becomes obsolete[30]

Digital preservation plans can be brief (1 to 2 pages), concisely communicating these objectives through one- or two-sentence statements. Many institutions make their digital preservation plans available online, and each is unique to the organization, its structure, and its goals. These online publications are invaluable resources for institutions establishing new preservation programs. Understanding how others address their commitment, needs, and resource allocation can inform decision making at your own institution. The good news is that you do not have to write a digital preservation plan from scratch; when you find one that you feel meets your needs, consider adapting it to your local conditions.

Preservation Policies

A digital preservation policy is the framework around which a digital preservation strategy is developed at an institution. It is integral to documenting the institution's commitment to digital preservation services, identifying at a high level what digital content is in scope of the digital preservation policy (in this case, digital audiovisual collections) and providing a vision for moving these strategies into action. As noted above, many institutions make their digital preservation policies available online, and templates are available as well.

A digital preservation policy document may include the following types of information:

- Mission and vision statements
- Definitions of terms
- References to other policies (collections, preservation of analog collections, etc.)
- Preservation strategy
- Process workflows
- Definitions of preservation levels (bit-level preservation versus full preservation of the files as they exist today)
- Definitions of acceptable preservation formats and metadata standards
- Overview of storage and technical infrastructure
- Definitions of communities (producers of content, users of content, etc.)

Policies and plans should be approved by high-level administration to ensure that administrators understand the goals of a digital preservation program and to solidify their ongoing support to sustain the program. Policies also facilitate buy-in from colleagues and staff as they integrate the digital preservation program as a key function of the organization.

Resources

Plans and Policies: Samples and Templates

- Bishoff, Liz. "Digital Preservation Plan: Ensuring long term access and authenticity of digital collections." *Information Standards Quarterly* 22:2, 2010. *https://groups.niso.org/apps/group_public/download.php/4250/FE_Bishoff_Digital_Preservation_Plan_isqv22no2.pdf*
- Dartmouth College Library Digital Preservation Policy, 2015. *https://www.dartmouth.edu/~library/preservation/docs/dartmouth_digital_preservation_policy.pdf*
- de Jong, Annemieke. *Digital Preservation Sound and Vision: Policy, Standards and Procedure,* 2016. *http://publications.beeldengeluid.nl/pub/388*
- MetaArchive Cooperative. "Digital Preservation Policy Template." 2010. *https://metaarchive.org/wp-content/uploads/2017/03/ma_dp_policy_template.pdf*
- NEDCC. "NEDCC Digital Preservation Policy Template." *https://www.nedcc.org/assets/media/documents/SoDAExerciseToolkit.pdf*

30 Liz Bishoff, "Digital Preservation Plan: Ensuring long term access and authenticity of digital collections." *https://groups.niso.org/apps/group_public/download.php/4250/FE_Bishoff_Digital_Preservation_Plan_isqv22no2.pdf*

SECTION 3: STORAGE INFRASTRUCTURE

In the digital environment, how digital content is stored is paramount. As noted above, standards such as the *ISO 14721: Open Archival Information System* (OAIS) and *ISO 16363: Audit and Certification of Trustworthy Digital Repositories* (TDR) provide a framework for how repositories should be structured and managed and what actions should be taken on the digital content within them. For example, best practice suggests that digital content be stored on "active" servers that are backed up and managed with preservation in mind. Storage on fixed devices, such as DVDs or external hard drives that are not monitored or backed up, do not meet the requirements of a TDR. And, they have high failure rates—the bits rot and the media inevitably degrades, which can mean catastrophic loss to collections stored on them.[31]

The often-heard "storage is cheap" aphorism is true when it comes to per-bit storage costs at scale. However, the reality for many institutions, especially small and mid-sized organizations, is different. The cost of digital collections storage can be significant, especially for audiovisual collections that are terabytes or petabytes in size. For this reason, cost often plays a significant role when an institution selects the type of storage and determines how it is managed. The good news is that there are less-expensive options that still allow for an acceptable level of maintenance of digital collections. Understanding what is available, the associated costs, and the risks involved in the various options is key to choosing the best solution for an institution.

At a high level, storage options can be broken down into two categories: local or on-premise storage, and cloud or outsourced storage. Based on an institution's requirements, technical infrastructure, and resources, one or both options may be feasible. Decisions about what type of storage works best for an institution's needs should be influenced by such factors 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.*** Who will take responsibility for managing the digital content—digital collections managers only, the entire archives staff, or someone else? Will a version of the content also be publicly accessible?
- ***Types and amount of digital content.*** How much storage do you need? At what rate will it grow?
- ***Redundancy.*** Is an institution capable of safely managing two or more copies of its digital content locally, or must it rely on cloud storage?

Considering these issues alongside best practices such as those in the NDSA Levels of Digital Preservation (see "Section 1.5: Resources"), levels of effort required, and the resources in place to support them will help an institution identify the best storage options for its situation.

In the rest of this section, **storage media**, **storage architectures**, and **storage capacity** are detailed in an effort to provide practical guidance on best practices.

Storage Architectures

Online, nearline, and offline are terms 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.

Online: In this context, online 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 common. Examples of online storage include flash and spinning disk, both described below in "Section 3.2: Storage Media."

Nearline: In this case, digital content is available to users with some lag time, which can be a few seconds to a minute 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. (Magnetic tape is described below in "Section 3.2: Storage Media.") This tends to be

31 "Seagate slapped with a class action lawsuit over hard drive failure rates," PCWorld.com, February 2, 2016. *http://www.pcworld.com/article/3028981/storage/seagate-slapped-with-a-class-action-lawsuit-over-hard-drive-failure-rates.html*

an option used by larger institutions that have 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 in order to access the data on it. The most common offline media type used for digital preservation is magnetic tape. Offline storage is often used to backup digital content for long periods of time. This kind of storage architecture is cost effective, but it takes time to access because it is not connected to a network. For digital preservation, offline storage is often used for the third-copy backup, or disaster recovery copy, of digital content.

Why use one architecture over another for audiovisual content? There are a number of reasons, including the following:

- ***Cost.*** Offline storage tends to be the cheapest, but it 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. However, 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 Linear Tape Open (LTO) tape because they will not be accessed regularly and because **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 because it requires transfer over the internet, meaning that the available bandwidth in and out of your facility provides additional constraints.
- ***Scale.*** The scale of audiovisual 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 using online or nearline architectures, other copies can be stored offline (e.g. on LTO tape), which tends to be the most cost-effective method of storage.

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

Storage Media

There are a variety of storage media options available to digital collections managers. Some are widely accepted as preservation-appropriate, while others are recognized as problematic due to their susceptibility to failure and obsolescence.

Removable media, such as portable hard drives, portable flash drives, or CDs and DVDs, are not considered viable as part of an overall preservation strategy. These media are highly susceptible to failure from degradation of the components that comprise them. The reality is that files are often stored on or burned to them and then the media itself is filed away or stored in boxes and not actively monitored. These media are also resource intensive to monitor for errors; human intervention is required to plug in a flash drive or play DVDs to identify errors in the files stored on them. For audio or video collections, this can mean hundreds or more of DVDs to monitor. If errors are identified on a DVD, it is typically the entire media object that has

FIGURE 4.2

Examples of removable storage media in use from 1970 to 2010, including 8" floppy disk, 5.25" floppy disk, 3.5" floppy disk, cassette tape, 8mm tape, CD, DVD, ZX Microdrive, SDHC card, CompactFlash card, and USB disk.

Credit: https://commons.wikimedia.org/wiki/File:Forty_years_of_Removable_Storage.jpg

failed, which leads to complete data loss or cost-intensive data recovery procedures that are out of reach for some institutions.

Two predominant storage media that are considered good options for digital preservation management are referred to colloquially as "spinning disks" and "magnetic tape." This is by no means an exhaustive list, but it provides some idea of the options available for the purposes of managing digital content.

Spinning disk storage that is part of a networked storage environment is commonly used in digital preservation environments. Spinning disk storage has quick response times and allows for active monitoring, such as fixity checking, to take place. This type of storage is often highest in cost because the media is expensive, the servers are always on and must be maintained in an environmentally controlled and secure area, and staff must be available to keep the servers up and running.

Magnetic data tape, most commonly LTO tape, is typically used either for nearline or offline storage (described in "Section 3.1: Storage Architectures"). Magnetic tape media is less expensive than spinning disk or other low latency storage options, and the cost of managing it over time is greatly reduced, especially for offline storage. Like other removable media, its mediated nature slows access and preservation activities such as active fixity monitoring. However, magnetic tape is much more reliable and far less prone to failure than portable drives and optical disc media such as CDs and DVDs. Magnetic data tape can be stored in what is known as a tape robot, which can provide some automation for access and preservation activities.

FIGURE 4.3
The inside of a spinning disk hard drive. Note how the head extends over the disk similar to the way a stylus rests on top of a vinyl record.
Credit: AVP

FIGURE 4.4
An LTO data tape cartridge.
Credit: https://commons.wikimedia.org/wiki/File:LTO2-cart-purple.jpg

Media that enables digital collections managers to actively monitor the health of their collections is always the best choice when deciding on storage options. Luckily, this type of storage also tends to be the most prevalent. No matter what choice you select for storage, though, always backup your data at least once and ideally twice (three copies total), or more.

Options for Storing Digital Collections

A major factor in deciding what type of storage to choose for digital content management, beyond what your institution already has in place, is cost. And when determining cost, it is important to take all of the costs of managing storage into account. The total cost of ownership (TCO) considers all of the media, labor, and overhead costs that go into installation, ongoing management, and even migration from one storage option to another at some point in the future. Of all the costs, ongoing management is the highest, so institutions more frequently consider cloud storage as a way to alleviate the day-to-day costs and responsibilities for storing digital content. Whether cloud storage is the answer depends on each institution's local organizational, resource, 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 storage locally (i.e. "on premise"). Digital collections managers should develop strong relationships with IT staff who manage storage at their institution, so that they can work together to build the best local 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, Amazon Glacier, and Google Cloud 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 are:

- **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 backed up to multiple locations?
- **Security.** What services are in place to ensure your data is safe?
- **Disaster recovery.** What happens if systems fail?
- **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?

COMPARISON OF CLOUD AND LOCAL STORAGE

Before deciding on one solution over another, a comparison of the features of each, in relation to the need for long-term management of digital collections, should be undertaken. Some considerations are listed in the accompanying table.

Each type of storage has its own financial and organizational implications, and each institution will need to weigh the factors above to come up with a solution that best suits their needs. In some cases, it will not be an either/or decision but a solution that uses both types of storage to their best effect for the institution's unique situation.

For example, one institution might have a mandate to maintain all collections, whether digital or physical, onsite. In this case, they may opt for a local-only storage solution. Another institution might not have the infrastructure and staff to manage collections onsite, due to costs or personnel restrictions, and may opt instead for cloud storage (from Amazon, Microsoft, Google, etc.) or even a provider like Preservica or DuraCloud that offers a set of preservation services in addition to cloud storage. And, as is more and more often the case, an organization might opt for a hybrid approach. In this case, they may choose to keep a single online copy on local storage so they have quick access to files when they need them. Secondary and tertiary copies may be stored locally on online or nearline storage or in the cloud. Often, yet another copy is stored on magnetic data tape (such as LTOs) in a different geographic location. These second and third backup copies tend to be versions of files that do not need to be accessed readily except for periodic fixity checks. This hybrid approach is an excellent way of (a) alleviating single points of technology failure by distributing content across storage solutions and (b) distributing content across geographically diverse locations.

TABLE 4.1

Comparison of cloud and local storage.

Credit: AVP

	Cloud	Local
Cost	Change in cost of services over time are unknown. Pricing is frequently akin to early cell phone plans; i.e. there are lots of unknowns until you're "in." Pay as you go. Only pay for what you use. Amount of admin required is typically unknown up front.	Most storage systems last 5–7 years. The cost of replacement must be taken into consideration. A significant portion of costs are upfront to pay for new technology, although ongoing costs for staffing and facilities is also a factor.
Staffing	Requires some staff to configure options, troubleshoot with technical support, and coordinate efforts. This is less of a staffing burden than local storage.	Requires dedicated staffing to manage infrastructure and users.
Support	Support will be different depending on the service provider. Because the user bases tend to be large, generalized services such as knowledge bases or FAQ pages are available.	Support is dependent on the IT staff responsible for managing the storage environment.
Exit Path	Many cloud storage plans make it cheap to upload content but very expensive to download it. Different cost models exist, and it is important to consider them carefully.	There is a clear exit path that is straight forward, although it requires more logistical planning and coordination on the part of the IT staff and the digital collections manager.
Scalability	It is relatively easy to increase the amount of storage you need—it is cost dependent.	Typically, scalable but may take more staff time and financial and computing resources to grow storage capacity.
Forward Looking	Storage and computing in general are trending toward the cloud.	It may pay off to take a "wait and see" approach with cloud storage, so you have more time to understand the true nature of cloud storage and computing as it matures and is tested over time.
Sustainability	Pay-as-you-go provides for more predictable financial planning over a longer period of time but requires continual investment. If funding stops or goes away, there are few/no options for what to do with the stored content.	Ongoing funding is required, and when technology must be updated, there are short-term capital cost increases. If funding stops or goes away suddenly, the infrastructure exists to buy time while you come up with alternatives.

Resources

General

- Griffith, Eric. "What Is Cloud Computing?" *PCMag.com*, May 3, 2016. *http://www.pcmag.com/article2/0,2817,2372163,00.asp*
- "Nearline storage." Wikipedia. *https://en.wikipedia.org/wiki/Nearline_storage*
- Newman, Jared. "Seagate slapped with a class action lawsuit over hard drive failure rates." *PCWorld.com*, February 2, 2016. *http://www.pcworld.com/article/3028981/storage/seagate-slapped-with-a-class-action-lawsuit-over-hard-drive-failure-rates.html*
- "Storage." *Digital Preservation Handbook*. *https://www.dpconline.org/handbook/organisational-activities/storage*

Standards

- ISO 14721:2012, Space data and information transfer systems—Open archival information system (OAIS)—Reference model. *http://public.ccsds.org/publications/archive/650x0m2.pdf*

- ISO 16363:2012, Space data and information transfer systems—Audit and certification of trustworthy digital repositories. *https://public.ccsds.org/pubs/652x0m1.pdf*

SECTION 4: ACTIVE MANAGEMENT

Digital preservation requires active management to ensure that digital assets persist over time. Compared to analog counterparts that may need little beyond stable environmental conditions, digital content requires constant monitoring for changes and mitigation should changes arise.

Change can occur in a variety of ways. Files may be updated or altered intentionally for good or for nefarious purposes. Natural disaster, hardware and software failure, **bit rot**, and human error can all affect the stability of digital content, as well. A key component of active management is an awareness of the threats posed by change and employing systems to reduce the chances that change can occur.

This section describes some of the key functions needed to actively and fully manage digital content for preservation.

Redundancy and Geographic Separation

Maintaining multiple copies of digital content in different geographic locations is a fundamental practice of preservation, whether in the physical domain (more than one institution may preserve copies of the same film) or in the digital domain (the same audio file may be stored on servers in Los Angeles and New York). It also means ensuring that your digital content is stored on more than one type of media; for example, spinning disk and data tape.

In the digital realm, it is ideal to maintain three copies of your digital content, stored in different geographic locations, on different types of media, and maintained in such a way that the copies are always the same.[32] This ensures that if something happens to one copy in one location or on one type of media, at least two other unchanged copies of the digital content persist.

PRESERVATION DATA BACKUPS

Most of us have accidentally deleted a file that our IT department was able to restore, either in its entirety or in an earlier state based on when they last performed a backup of the server on which the file was stored. Active data backups involve copying actively used production files, often on a daily basis, for the purpose of short-term retention while the files are in use. These backups might be saved for a week or a month, but after a period of time they are overwritten by new backups. Access copies of digital audiovisual content—the files that you use and share on a frequent basis—are typically backed up in this way.

Preservation data backups are slightly different. Instead of files in active use, the best-quality versions of the digital files that are in a finished, inactive state (often referred to as "preservation masters") are copied from a primary storage location to a secondary (and ideally, tertiary) storage location. All three copies of replicated data are typically composed of identical "packages" that contain the digitized preservation masters, as well as preservation metadata to help identify and use the files when, for example, the access copies are no longer viable and must be replaced. The preservation masters are not meant to be accessed in the near term, however. The length of time that these preservation packages are backed up is dependent on their retention requirements. In the case of digital audiovisual content, the retention schedule is often "for as long as possible."

GEOGRAPHIC DISTRIBUTION

Ensure that copies are in geographically disparate storage locations/systems to decrease the likelihood of loss due to localized disaster or service interruption. For example, if your data center is located along the coastal United States, it is best to store duplicate copies of your digital content on servers in another region where hurricanes are not a concern. Or, for example, a university may store digital content on one type of media in campus buildings, on premise, and on another type of media in a satellite location, off site, 20 or 30 miles away. It may also maintain a third copy of the digital content even further away to provide geographic diversity, or even in the cloud (see Section 3).

32 Megan Phillips, et al. "The NDSA Levels of Digital Preservation: An Explanation and Uses," *www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_Levels_Archiving_2013.pdf*

Ideally, storage devices will allow preservation packages to be actively monitored so that errors in data can be identified and repaired. How that identification and repair happens is described in the next section.

File Fixity and Data Integrity

In the context of digital preservation, fixity describes the unchanged state or "fixed-ness" of a digital file. Fixity monitoring can identify whether a file has changed for any number of reasons, such as human error, hardware failure, or bit rot, so that action can be taken to repair it.

ZEROS AND ONES

Digital files are, at their most basic level, made up of a series of zeros and ones.[33] Each number is a "bit." A file is essentially a very long list of bits stored in a particular order that is different from the order of every other digital file, and this enables a computer to access and play it. A small change, such as turning a single zero into a one, changes the makeup of the entire file, sometimes in catastrophic ways.

Keeping track of all of the bits is hard. Files can be thousands or millions of zeros and ones long, and it is important to ensure that files are stable, or "fixed," so that they can be accessed. Instead of keeping track of the bits themselves, we can check the files' fixity by watching or listening to each of them on a regular basis, but that is time consuming and unrealistic in large collections. Luckily, there is another method for monitoring fixity to ensure that if a file does change, we catch that change and can repair it.

MONITORING FIXITY WITH CHECKSUMS

To make monitoring files for changes to the bits easier than keeping track of millions of zeros and ones, we use shorter, alphanumeric strings that reflect the uniqueness of every digital file. These strings are called checksums and are generated by a program that reads the zeros and ones of a file and creates a unique string of characters to represent them. This checksum becomes that file's signature and will remain the same as long as the bits do not change.

If a file is changed, even in seemingly insignificant ways (you may not even be able to tell just by playing a file), a completely different checksum will be produced by the checksum generator.

Checksums are valuable for several reasons. For example, they can be used to authenticate a file. If a file is the official version of a video, it can be authenticated by first creating a checksum signature and then running the checksum program later to be sure that the signature has not changed. Or, if a file is being deposited in an archive, a checksum may be produced before deposit and then after deposit. This one-time test authenticates that the file is what the depositor understands it to be and that it was not corrupted upon ingest.

One of the greatest values of checksums is their use in monitoring file fixity, which tests for fixedness or stability of the bits over time. It is important to note that while different files have different checksum signatures, exact copies of files will have the same signature. As long as a file doesn't change, it will always have the same checksum signature as other identical copies. That means files can

FIGURE 4.5

Image of a list of checksums, whose signatures are on the left, and their associated files. The algorithm used to create the checksums in this example is called SHA256, although there are others, such as MD5 and SHA1, in wide use as well.
Credit: AVP

```
99 00:00:00 99 99
2016-08-18 15:43:57
||-||0
sha256
3305a6ec078c2627991e86b7385459dbf43e983bee35de6cc13dd7c26e44c5ff    /Users/.../photos/choc_pudding.jpg    93656899
218a3e8add730d479bb492c3ef86f1c10084e311c2c79267c0f706b48c33ed8b    /Users/.../photos/green_beans.jpg     93656846
24bd9d27a52a11c5edd4a4bc7d75edbb2724cecbf44541e0ebfcb7c34df0484a    /Users/.../photos/ice_cream.png       93656879
f55a999e834bd9c6660bbfd676f65860ebefb2d9f008ad5469a1f311cb631653    /Users/.../photos/potato_salad.jpg    93656713
```

33 Computers use binary code (zeros and ones) to process information.

be monitored for change using a tool that checks fixity on an ongoing basis and repairs files when checksums do not match by replacing them with another unchanged copy of that file. Monitoring fixity over time (e.g. once every month, six months, or a year) allows you to identify changes and replace corrupt files with an unchanged copy.

Finally, while checksums are the primary mechanism for monitoring fixity at the bit level, they can also be used to monitor file attendance, or identifying if a file is new (the checksum signature has never been produced before), removed (a checksum is missing from a list), or moved (the checksum appears with files in another location). Tracking and reporting on file attendance is a fundamental component of digital collection management and fixity.

THE FIXITY TOOL

Some institutions have sophisticated systems and workflows that automate the monitoring of fixity and file attendance in their digital collections; however, many do not. A good place to start with fixity and tracking checksum signatures, if technology systems are not available to automate this process, is by maintaining an inventory in spreadsheet software that lists the files in your collections, their locations on your servers, and their associated checksums and the dates that they were produced. Over time, the inventory will help you identify changes at both the bit and file level so that you can find and repair the errors in your collections.

There are a variety of tools for creating and verifying checksums. Some of these include:

- ExactFile (Windows, *http://www.exactfile.com*). Calculates a variety of checksums.
- FastSum (Windows, *http://www.fastsum.com*). Calculates MD5 checksums.
- HashMyFiles (Windows, *http://www.nirsoft.net/utils/hash_my_files.html*). Calculates SHA1 and MD5 checksums.

In addition, all operating systems have built-in checksum generation and validation functionality; however, using them requires access via a command-line user interface.

An institution might also consider the free and open-source tool Fixity.[34] Fixity was created with smaller and/or lesser-resourced organizations in mind and is a simple application that enables automated checksum production and file attendance monitoring and reporting. Fixity scans a folder or file directory and creates a manifest of the files including their file paths and checksums, against which a regular comparative analysis can be run. Fixity monitors file integrity through generation and validation of checksums and file attendance through monitoring and reporting on new, missing, moved, and renamed files. Fixity emails a report to the user that documents flagged items along with the reason for a flag, such as that a file has been moved to a new location in the directory, has been edited, or has failed a checksum comparison for other reasons.

Information security

The ultimate goal of preservation is to ensure that collections remain unchanged and accessible over time. Information security protocols help to minimize accidental or nefarious changes by users and the public and to track how files are changed, who changed them, and when the alterations were made. Protocols help ensure that authenticity is maintained, and when it isn't, that a change is documented so that an organization can act on that change by, for example, replacing corrupted files with backup copies. This section describes some of the ways that information security is used to maintain the authenticity of digital collections.

USER PERMISSIONS

In a digital preservation environment, organizations must control which users are accessing and manipulating data. Some users may have access to view files, while others may have controls over where files reside, their formats, and who can access them. Creating, assigning, logging, and managing permissions and restrictions are critical to mitigating the risk of intentional or unintentional data corruption and misuse of content. Many preservation management systems make permissions management easy. However, when data management happens manually or outside of a management system, then setting access permissions on directories on networked servers can be a

34 *http://www.avpreserve.com/tools/fixity*

TABLE 4.2

An example of how digital preservation access permissions can be applied to users within an organization. Permissions will differ for each organization, and it is important that staff responsible for digital preservation have a voice in how they are applied.

Credit: AVP

Staff position	Access Permission on Storage Device			
	Read	Write	Move	Delete
Digital preservation staff	X	X	X	X
IT support staff	X	X	X	
Other internal users	X			

TABLE 4.3

An example of a manually constructed audit trail that includes user, file, action, and date.

Credit: AVP

User	File	Action	Date
Jane Smith	f10d6b9e.wav	Moved file from Directory 1 to Directory 2	2016-09-07

good way to create a secure space to store digital collections. IT staff can usually help implement these strategies.

Computer systems offer varying levels of permissions that enable or restrict access to digital collections. Four types of permissions are read, write, move, and delete. These functions are more or less what they sound like, although the concepts of "read" and "write" may not be entirely obvious. In this context, "read" access refers to the ability to view files in a directory without being able to edit or delete the files. "Write" access means that a user can add files to a directory or edit files within it. The levels of permissions are cascading, starting with the lowest level of access—none—and ending with the greatest–"delete." This means that a user with "delete" access also has "move," "write," and "read" access. Conversely, a user may have only "read" access or no access at all. If you are able to set permissions on directories that contain preservation copies of digital files, then consider doing so, but with caution. Always be sure that an administrator has full access to the collection, so that if the digital content needs to move, migrate, or be monitored for fixity, there are no restrictions on doing so.

As with all digital preservation activities, it is important to document decisions about permissions. It is one thing to assign permissions, but having information at hand that documents who has been assigned which permissions will help identify whom to contact when diagnosing data errors. Logging access and internal actions taken on digital collections is equally important, although challenging, in an environment where the logging is not an automated process. Logs, or audit trails, enable digital collections administrators to audit actions taken and track changes to a user and date, which can be valuable when trying to understand when and where errors have occurred. Although by no means impossible, these activities can be time consuming in a manual workflow. An example of a manually constructed audit trail might include the following notations (Table 4.3).

The good news is that many preservation management systems automate this work. Whatever preservation tools you have at your disposal, always remember to do the best you can do now, so that you're ready to do more when the resources and infrastructure enable it.

PROTECTING FROM EXTERNAL THREATS

Every organization needs to be concerned not only with controlling access to digital collections from within, but also with protecting assets from external threats. This is particularly important for storage devices such as servers that are connected via networks and to the internet. Controlling access to these devices with good password and username practices is imperative. Passwords should be

unique, high-strength (upper and lowercase letters, numbers, and symbols), long (12–15 characters is recommended),[35] and changed routinely.

Operating systems, when not updated routinely, can be another potential risk to a network. For example, at the time of this writing in 2017, Windows XP runs on 7% of computers across the globe, although support for it from Microsoft officially ended in 2014.[36] This means that security patches, created in response to active virus threats and other nefarious software, are rarely produced. This puts these computers and the networks on which they run at risk of security breaches. Taking a proactive approach to keeping operating systems up to date decreases the chances of data breaches.

VIRUS SCANNING

Another important component of information security is monitoring digital collections for viruses and other corrupting malware. Virus scanning should be performed on files that are being brought into a digital preservation environment from external sources to avoid infecting existing files and systems. If you are performing digitization internally and have full control over the files being created, this is less of an issue. It becomes a greater concern when accepting files from external sources, such as donors or even other units within an organization. Virus scanning should be performed on all external data transfers into the environment and then routinely in the digital preservation environment as an added precautionary measure. One approach is to have a dedicated "clean" computer that is not connected to a network or the internet. Files can be virus tested on this machine before transferring them to networked storage. This may be especially useful to institutions where IT support is not readily available.

A good idea is to create a folder that is specifically used to store newly acquired files to prepare them for ingest into your preservation environment. The folder should not be directly connected to your preservation environment. If you cannot check the files on the media on which they are delivered, transfer new files directly into this folder and immediately perform virus checking on them. That way, if corruption is identified, the files cannot infect your existing digital collections.

There are many low-cost virus-scanning software options available on the market today. If you have an IT department, talk to them about virus scanning before you purchase your own software. They might have options available that are already in use in your organization.

Resources

- AVPreserve. "Fixity." *https://www.avpreserve.com/tools/fixity*
- De Stefano, Paula, et al. "What is Fixity, and When Should I be Checking It?" *Checking your Digital Content: An NDSA Publication*, 2014. *http://www.digitalpreservation.gov/documents/NDSA-Fixity-Guidance-Report-final100214.pdf*
- "Information Security." *Digital Preservation Handbook. https://www.dpconline.org/handbook/technical-solutions-and-tools/information-security*
- "Storage." *Digital Preservation Handbook. https://www.dpconline.org/handbook/organisational-activities/storage*

SECTION 5: METADATA

We recommend that you read this section together with "Metadata," in Chapter 3 (page 49), to get a complete understanding of the types of metadata pertinent to audiovisual collections, how metadata is generated, and standards for capture.

Metadata is the information about a digital file that allows us to understand, use, manage, and preserve it. Without it, we would not know about the file (e.g. the title, who created it, and on what date), what the file is (e.g. the wrapper or codec in use, data rate, pixel dimensions,

35 Brian Barrett, "7 Password Experts on How to Lock Down your Online Security." Wired.com, May 5, 2016, *https://www.wired.com/2016/05/password-tips-experts.*

36 Nicole Becher, "Patching is Hard, and Windows XP Will Die Harder," *http://www.slate.com/articles/technology/future_tense/2017/05/patching_is_hard_and_windows_xp_will_die_harder.html*

duration, etc.), how it relates to other files (e.g. part one of three), and how it has been monitored over its lifetime (e.g. fixity checks). Without the appropriate metadata, a file becomes inaccessible and unusable, ultimately losing its value.

Metadata is produced at various times during a file's lifespan. Descriptive and technical metadata are often captured at the point of creation, such as during the digitization process. Preservation metadata, on the other hand, is generated on an ongoing basis to log actions performed on the file, including such activities as transferring it from one storage environment to another, performing fixity checks, or migrating from one format to another. Logging preservation metadata over time creates an audit trail that ensures a file remains authentic and accessible for the long-term.

Types of Metadata

Each type of metadata plays a different role that, along with a digital file, makes up the information package that ensures long-term access. It is useful, then, to describe the different types of metadata and how they affect a digital file.

DESCRIPTIVE METADATA

Descriptive metadata is the information about a file or files that enables identification and discovery. Descriptive metadata includes the title, creator, date of creation, and keywords that document the subject of the file's content.

STRUCTURAL METADATA

Structural metadata is the information that designates how a set of files relate to one another, such as songs on a CD or how the parts of a single file are structured. For example, it could tell you that a file comprises a container file with one video, two audio, and two subtitles tracks.

ADMINISTRATIVE METADATA

Administrative metadata includes information about how to manage a digital file and track its process history. This ranges from rights metadata, which indicates who owns or holds copyright for a file and how it can be used and accessed, to **technical** and **preservation metadata**, which are described in detail below.

While all forms of metadata help provide long-term access to digital collections, technical metadata has particular significance for audiovisual content and preservation metadata is key to ensuring that digital content can be managed over time.

TECHNICAL METADATA

Technical metadata captures the essence of a digital file. It is the technical information that describes how a file functions and that enables a computer to understand it at the bit level, so that it can be played back in a way that is useful for a viewer or listener. Technical metadata includes information such as wrapper, codec, compression, and aspect ratios. Often, technical information is embedded in the file itself and is read directly by compatible software and hardware. For preservation purposes, technical metadata is extracted and stored outside the file so that as formats obsolesce and compatibility fades, a file's basic structure is understood and can be migrated to a new format that allows it to be accessed. Many audiovisual metadata schemes rely on technical metadata—extracted from embedded metadata—as one method for maintaining the usability of digital content over time.

PRESERVATION METADATA

Preservation metadata is the information necessary to support the management and long-term accessibility and usability of an object. It tracks the processes that are necessary to manage a file in a digital environment over time, including: monitoring fixity and performing any repairs that are identified during fixity checks, auditing logs to identify when and who has interacted with an object, monitoring obsolescence information, and documenting provenance information to support the authenticity of an object. Examples of preservation metadata include checksums, storage locations, and records of process activities and dates (for example, that a file is moved from one location to another and the date that the move occurred).

Embedded Metadata

Embedded metadata is the information that is stored within a file that also stores the content to which the metadata refers. For example, a WAVE file contains both the music and the technical information to play the file. Embedded metadata can also include descriptive information,

as in mp3 files, which enables display of an artist, album, and title in applications that play them.[37] Thinking about it another way, embedded metadata is the digital equivalent of physical labels, annotations, and written documentation stored inside a material housing or the video slates at the head of a recording.

Embedding information about the holding organization (the data source that holds information about the object) and the copyright status also helps to identify the file if it becomes disassociated with the metadata that is part of its information package. The Federal Agencies Digitization Guidelines Initiative (FADGI) is a set of published guidelines for digitization processes and offers guidance for the use of embedded metadata in WAVE files.[38] For example, the guidelines offer recommendations about how to store embedded metadata in WAVE files that result from the digitization process.

EXTRACTION TOOLS

Because of the embedded nature of much technical metadata, tools have been developed to automate the extraction of this information from the files in which it is held. Two of those are:

- **FITS,** *https://projects.iq.harvard.edu/fits*, is a command-line tool developed by Harvard University Library that identifies, validates, and extracts technical metadata for digital objects, including some audiovisual formats. The metadata is exported into an XML file.
- **MediaInfo**, *https://mediaarea.net/en/MediaInfo*, is an open-source program that extracts technical metadata about media assets and exports it into a variety of formats including txt, EBUCore, PBCore, and reVTMD, which are all described below. It works with a variety of audio and video formats and has a GUI interface, so command-line knowledge is not necessary. It is available for many operating systems.

Standards, Schemas, and Guidelines

Metadata standards, schema, and guidelines are invaluable to the creation, management, and sharing of information. They tell us how and why certain metadata should be captured, enabling us to easily understand metadata created by others and minimizing the obstacles of sharing information between systems. Metadata can be stored in Excel spreadsheets, as XML files, or in databases such as content management systems and institutional repositories, as well as in other formats. However metadata is stored, using standards to create and structure it will make it more broadly understood and interoperable.

The standards and guidelines briefly described below are just a few of the most recognized and recommended for the management of audiovisual collections.

EBUCORE

EBUCore is based on the Dublin Core standard and adapted to broadcast media. It is a descriptive and technical metadata schema developed and maintained by EBU, the largest professional association of broadcasters in the world. EBUCore captures the minimum information needed to describe radio and television content.

Link to more information about EBUCore:
https://tech.ebu.ch/docs/tech/tech3293.pdf

Link to the EBUCore metadata specification:
https://tech.ebu.ch/MetadataEbuCore

FADGI (FEDERAL AGENCIES DIGITIZATION GUIDELINES INITIATIVE)

Begun in 2007, this is a collaborative effort by US federal agencies to define common technical guidelines, methods, and practices for digitizing historical content and the capture of technical metadata. The focus of the audiovisual working group, in particular, is to identify, establish, and disseminate information about standards and practices for the digital reformatting of historical and cultural audiovisual materials by federal agencies, although the guidelines

37 *http://id3.org*

38 Federal Agencies Audio-Visual Working Group, "Embedding Metadata in Digital Audio Files: Guideline for Federal Agency use of Broadcast WAVE Files," version 2. *http://www.digitizationguidelines.gov/audio-visual/documents/Embed_Guideline_20120423.pdf*

have seen broad use beyond the US government as well. The effort covers sound recordings, video recordings, motion picture film, and born-digital content.

Link to more information about the FADGI audiovisual working group and its guidelines: *http://www.digitization-guidelines.gov/audio-visual*

METS (METADATA ENCODING AND TRANSMISSION STANDARD)

The METS schema is a standard for encoding descriptive, administrative, and structural metadata about objects within a digital library, expressed using XML. METS provides an XML document format for encoding metadata necessary for both the management of digital library objects within a repository and the exchange of such objects between repositories or between repositories and their users. METS is a Digital Library Federation initiative that is maintained by the Library of Congress.

Link to more information about METS: *https://www.loc.gov/standards/mets/mets-home.html*

PBCORE (PUBLIC BROADCASTING METADATA DICTIONARY PROJECT)

PBCore is a metadata schema designed for sound and moving images. It can be used as a guideline for cataloging the descriptive and administrative information about audiovisual content. It can also act as an exchange mechanism to share information between institutions or applications, and much more. PBCore expands on the Dublin Core standard. It was created by the US public broadcasting community and is maintained by WGBH in Boston.

Link to more information about PBCore: *http://pbcore.org*

PREMIS (PRESERVATION METADATA: IMPLEMENTATION STRATEGIES)

The PREMIS Data Dictionary for Preservation Metadata is the international standard for metadata to support the preservation of digital objects and ensure their long-term usability. PREMIS is a comprehensive, practical resource for implementing preservation metadata in digital archiving systems.[39] It is maintained as a standard by the Library of Congress.

Link to more information about PREMIS: *http://www.loc.gov/standards/premis*

REVTMD

reVTMD is an XML schema tailored to include fields that address the creation and long-term management of reformatted videos, especially for the cultural heritage community. It is a concise subset of the large array of technical metadata available for digital media, structured in a way to make it highly usable for accessing and managing all types of video files. The captureHistory section is especially helpful in capturing process history for preservation purposes. reVTMD was developed by the US National Archives and Records Administration in collaboration with AVPreserve.

Link to more information about reVTMD: *https://www.weareavp.com/tag/revtmd*

Link to the reVTMD XML schema: *https://www.archives.gov/preservation/products/reVTMD.xsd*

Resources

- Baca, Murtha, ed. *Introduction to Metadata*, 3rd edition. *http://www.getty.edu/publications/intrometadata*
- Greisinger, Peggy. "A brief overview of metadata for audiovisual materials," November 6, 2014. *http://ndsr.nycdigital.org/a-brief-overview-of-metadata-for-audiovisual-materials*
- International Association of Sound and Audiovisual Archives. "Metadata." In *Task Force to establish Selection Criteria. http://www.iasa-web.org/task-force/7-metadata*
- Zeng, Marcia L. *Metadata Basics. http://metadataetc.org/metadatabasics*

39 Library of Congress, PREMIS Data Dictionary for Preservation Metadata, *http://www.loc.gov/standards/premis*

SECTION 6: PLANNING FOR OBSOLESCENCE

As this chapter has made clear, managing digital content requires a great deal of planning to ensure that the three-legged stool of organization, resources, and technology remains balanced over time. And, like so much related to digital preservation, this planning cannot stop; it is a long-term commitment to long-term access of digital content. Planning today takes into consideration the needs of tomorrow to ensure that wrappers, codecs, and the media on which digital objects are stored do not obsolesce–and if they do, that there are established pathways for moving them to new formats and media so that they remain accessible over time. (For more information on reformatting, see Chapter 3.) It's not only formats that obsolesce but also the media on which they are stored and the systems that are required to read or play them. If, for example, an mp3 music file is on a CD ROM but you no longer have a player with which to listen to it, the viability of the file format is only one concern; finding a way to read the media is another complicating factor.

Obsolescence monitoring

A fundamental practice of digital preservation is monitoring wrappers, codecs, and media to ensure they remain usable and that they have not become obsolete. Obsolescence can happen at the file or program level (such as Real Media and WordStar) or with the media on which the file is stored (such as Zip drives). When this happens, the digital content becomes unreadable or inaccessible.

Before obsolescence happens, steps can be taken to transition to current storage options or formats, such as Microsoft Word. Keeping track of changes in technology, called "obsolescence monitoring," involves maintaining awareness of file formats, software, and systems that are ubiquitous (which have less of a chance of becoming obsolete) and, more importantly, awareness of more specialized or less-used formats, which have a greater chance of obsolescence. Monitoring can be accomplished without specialized technologies simply by keeping up with the changing landscape; however, technology-based watch systems are available. Paired with proactive planning, these systems can help ensure that digital content does not become obsolete.

FIGURE 4.6
An LTO data tape cartridge.A screen shot of WordStar 4.0.
Credit: AVP

The good news is that obsolescence tends to happen slowly. Routinely reviewing files, conferring with colleagues, and learning about industry-standard formats are some of the best strategies for ongoing obsolescence monitoring.

Refreshing and Migration

The ultimate goal of digital preservation is the ongoing accessibility of digital content. Think of "content" in this context as you would the moving image on a reel of film. The physical object will change when the film is replaced by a reformatted version stored on digital media, but the content—the images and sound and the order in which they are played—remain the same. Digital audiovisual content must be able to transcend software and hardware changes over time, so choose non-proprietary or "open" formats when possible, such as Broadcast WAV for audio. The media on which digital files are stored will also become obsolete or unstable with age—usually more frequently than file formats do—and those files will need to move to more up-to-date media. Some estimates put the lifespan of hard disk drives at three to five years; others put their median lifespan at six years or more.[40] When the risk of loss of digital content due to the threat of obsolescence becomes too great, we look to refreshing and migration.

Refreshing refers to the approach of transferring files and/or data from one media, server, or system to another. This may consist of moving files from an aging server to a new one or shifting metadata from an obsolete database to a more widely used system. Great care must be taken to ensure that all data is transferred in a "lossless" way and that the integrity of the content is verified after the move. Error checking could include running fixity checks on files moved from one server to another, looking for changes or missing files, or reviewing a significant sample of metadata records (10% or more). The frequency with which refreshing happens depends on the technology; data on servers should be refreshed at least every three to five years. Shifts in metadata databases are dependent on the ongoing maintenance of the system in which they are stored.

Migration refers to the transfer of the content and metadata from one audiovisual format, such as a wrapper and/or codec, into another. Migration is necessary when the wrapper/codec or the software used to read the format becomes less ubiquitous and is on the verge of becoming inaccessible. Migration may consist of transferring audio from one wrapper to another without changing the codec, or it may consist of transcoding the audio and placing it in a new wrapper. Real Video, a proprietary video file and encoding format, used to be relatively common for video streaming on the web in the late 1990s and early 2000s. Today the ubiquity of formats like MP4 with h.264 encoding have rendered Real Video practically obsolete. If care is taken to ensure that video and audio are captured at high resolution using stable and open or uncompressed formats, the need to use migration for preservation will be less likely. However, migration, or perhaps the creation of new derivatives from preservation masters, might be required as web trends shift over time.

Because obsolescence can happen at many levels (codec→wrapper→content management system→storage media), refreshing and migration plans must consider all of these: hardware (e.g. servers), software (e.g. video platforms, codecs), and databases (e.g. digital asset management systems [DAMS], collection management systems [CMS]).

Initially selecting the codecs and wrappers, systems, and media with the greatest longevity and openness is ideal and means that refreshing and migration are not immediate concerns. While refreshing and migration are inevitable, being able to postpone that need means you can focus on other aspects of digital preservation management. That isn't to say that planning should not happen. Identifying funding that will be required when new hardware must be purchased or when staff are hired to complete the process of migration and refreshing helps to future-proof (and disaster proof) digital collections.

40 Brian Beach, "How long do disk drives last?" Nov. 12, 2013. *https://www.backblaze.com/blog/how-long-do-disk-drives-last*

Resources

- "Digital Preservation Strategies." In *Digital Preservation Management: Implementing Short-Term Strategies for Long-Term Solutions.* *https://dpworkshop.org/dpm-eng/terminology/strategies.html*
- "Obsolescence: File Formats and Software." In *Digital Preservation Management: Implementing Short-Term Strategies for Long-Term Solutions.* *https://dpworkshop.org/dpm-eng/oldmedia/index.html*
- *Sustainability of Digital Formats: Planning for Library of Congress Collections.* *http://www.digitalpreservation.gov/formats/intro/intro.shtml*

SECTION 7: PRIORITIZATION AND PHASING

Approaching digital preservation as a whole can be intimidating and overwhelming. Prioritizing approaches based on a digital preservation plan can assuage those feelings. Instead of thinking about the "foreverness" of digital preservation, consider it in periods of five- or 10-year increments. At the end of each period you have the option to make a decision about what to do in the following five or ten years. You can choose to do nothing, do the minimal amount necessary to maintain the option to decide again later, or you might choose to pursue a more robust solution, in whole or in part.

The aphorism "perfect is the enemy of good" is a useful way of thinking about prioritizing and phasing digital preservation activities. Doing something now—for example, creating an inventory of your collections, developing a collection policy, or making a backup copy of your digital content—is better than waiting for the perfect technological solution along with the resources and organizational framework to support it.

So, how to prioritize? Consider the following factors:

- **Impact.** Small steps can have great influence. Proper file naming, good metadata, and the use of open format wrappers and codecs are a few strategies that will have a significant impact on the longevity of your digital collections. They do not require complex systems and expensive technology to implement but they make digital content more findable, understandable, and usable over time.
- **Feasibility.** The reality is that some digital preservation practices require technology and resources that not all institutions can staff or fund. Other activities may be too time consuming for one person to undertake. Figure out what you can manage right now, and do it. As your capacity and experience grows, you might find that you gain support for more technologies and resources, making more complex activities more feasible.
- **Urgency.** An important consideration in making sure you are effectively using resources—especially if they are limited—is identifying what needs to be done right away. For example, moving digital content off of fragile or obsolescing media such as CDs and onto actively managed servers before data is lost may be of greatest urgency for your institution. Identify problems that, if you do not address them, will become more significant. Also, if opportunities arise, such as a one-time source of funding, be prepared to take advantage of them in a timely manner. Of all the factors, urgency should take priority, especially if digital content is at immediate risk.

Being flexible, putting what you know into practice, and taking a proactive approach today will establish a foundation that makes implementation and adoption of new technologies and programmatic preservation strategies possible and easier in the future.

NDSA Levels of Digital Preservation

Sometimes it is hard to know where to start. With so many standards and guidelines, beginning to address the challenge of long-term, active management of your digital content can be challenging. The National Digital Stewardship Alliance (NDSA) had this challenge in mind when it developed the "Levels of Digital Preservation."

The "Levels of Digital Preservation" are a tiered set of recommendations for how organizations can begin to build or enhance their digital preservation activities.[41] The Levels (LoDP) are meant to be an easy-to-use set of guidelines

41 Levels of Digital Preservation. *http://ndsa.org/activities/levels-of-digital-preservation*.

for those beginning to think about digital preservation, as well as for those with established programs that are ready to take the next step and enhance services. The focus of LoDP is on the content in digital collections and the infrastructure in place to manage it. LoDP is not designed to assess the overall readiness of digital preservation programs, like ISO 16363 does (see "Section 1.4: Standards"), because it specifically addresses technology and not issues related to organizational infrastructure.

The LoDP guidelines are organized into five functional areas that are at the heart of digital preservation technology systems and which are addressed in more detail in "Section 4: Active Management:"

- storage and geographic location
- file fixity and data integrity
- information security
- metadata
- file formats

TABLE 4.4

National Digital Stewardship Alliance's Levels of Digital Preservation (2013)

Credit: NDSA

	Level 1 (Protect your data)	Level 2 (Know your data)	Level 3 (Monitor your data)	Level 4 (Repair your data)
Storage and Geographic Location	• Two complete copies that are not collocated • For data on heterogeneous media (optical discs, hard drives, etc.) get the content off the medium and into your storage system	• At least three complete copies • At least one copy in a different geographic location • Document your storage system(s) and storage media and what you need to use them	• At least one copy in a geographic location with a different disaster threat • Obsolescence monitoring process for your storage system(s) and media	• At least three copies in geographic locations with different disaster threats • Have a comprehensive plan in place that will keep files and metadata on currently accessible media or systems
File Fixity and Data Integrity	• Check file fixity on ingest if it has been provided with the content • Create fixity info if it wasn't provided with the content	• Check fixity on all ingests • Use write-blockers when working with original media • Virus-check high risk content	• Check fixity of content at fixed intervals • Maintain logs of fixity info; supply audit on demand • Ability to detect corrupt data • Virus-check all content	• Check fixity of all content in response to specific events or activities • Ability to replace/repair corrupted data • Ensure no one person has write access to all copies
Information Security	• Identify who has read, write, move and delete authorization to individual files • Restrict who has those authorizations to individual files	• Document access restrictions for content	• Maintain logs of who performed what actions on files, including deletions and preservation actions	• Perform audit of logs
Metadata	• Inventory of content and its storage location • Ensure backup and non-collocation of inventory	• Store administrative metadata • Store transformative metadata and log events	• Store standard technical and descriptive metadata	• Store standard preservation metadata
File Formats	• When you can give input into the creation of digital files encourage use of a limited set of known open formats and codecs	• Inventory of file formats in use	• Monitor file format obsolescence issues	• Perform format migrations, emulation and similar activities as needed

Each functional area is gauged against a set of criteria that help an institution identify their own level of digital preservation readiness. The four levels are progressive; the requirements in the first level are building blocks for levels two through four. Although level one is the foundation on which the other levels are built, sometimes an institution's readiness in another level is higher than it is in level one.

The important thing, and what LoDP can offer to a new or growing digital preservation program, is the ability to prioritize the needs of an organization, effectively identifying what the organization is capable of attaining at the present moment. This means that the organization can be at a readiness level of three without having fulfilled all of the requirements of level two in the same category. After level one is addressed, the organization can shift its focus to filling in missing requirements in level two and progressing onward to level four. The move towards a programmatic solution should be continuous, fluid, and flexible with the understanding that a preservation program is being built that can withstand future organizational changes.

Resources

- Hodges, Patricia, et al. "The Five Organizational Stages of Digital Preservation." *Digital Libraries: A Vision for the 21st Century: A Festschrift in Honor of Wendy Lougee on the Occasion of her Departure from the University of Michigan*, 2003. *https://quod.lib.umich.edu/s/spobooks/bbv9812.0001.001/1:11/–digital-libraries-a-vision-for-the-21st-century?rgn=div1;view=fulltext*
- National Digital Stewardship Association (NDSA). "Levels of Digital Preservation." http://ndsa.org/documents/Levels_v1.pdf
- NEDCC. "Preservation Leaflet 1.4: Considerations for Prioritization." *https://www.nedcc.org/free-resources/preservation-leaflets/1.-planning-and-prioritizing/1.4-considerations-for-prioritizing*

CONCLUSION

Digital preservation is a multi-faceted endeavor that requires institutional planning and policies, a suitable storage infrastructure, active management, and well thought-out and consistently applied metadata. Careful planning and thoughtful policies will provide a program with concrete, achievable goals and will help to ensure the organizational commitment needed to make a program sustainable over time. A storage infrastructure that addresses the needs of the institution while accounting for the costs involved will promote sustainability as well, and the options described in this chapter will empower organizations to make a thoroughly informed decision. However, without vigilant monitoring and management, the security and fixity of files cannot be assured, no matter how well designed the storage is for digital collections. Metadata that complies with standards and is informed by guidelines will facilitate this monitoring and keep digital audiovisual collections accessible into the future. These components, combined with planned migration of collections to address obsolescence, constitute the makings of an effective digital preservation program for audiovisual collections.

This series of programmatic elements can at first appear daunting to those at repositories seeking to reformat and care for audiovisual media. Developing a sense of your institution's priorities will make phasing these components into a preservation program possible. The creation of these priorities should take into account their potential impact, feasibility, and urgency. The NDSA Levels of Digital Preservation can provide a helpful framework to further assist in this decision making.

Because a preservation program is so dependent on the unique context of an institution, there is no single correct way to create a program that is sustainable over time. Using the information outlined in this chapter will best prepare you to chart a course that meets your institution's specific needs at a pace that takes into account the resources available to you. In "Chapter 5: Disaster Preparedness and Response," you will learn about the considerations, strategies, and tips that will help you to secure and salvage your collections when faced with an emergency and to ensure that your developing preservation program is ready for the unexpected.

GLOSSARY

Active management: The performance of consistent and ongoing digital preservation activities (e.g. fixity and validation) to ensure a digital file's continued access for as long as necessary.

Artifact: Anomalies during visual or aural representations of recordings.

Audit trail: The information associated with a digital file that tracks the transactional history of it from the point of capture or ingest to know whether it has been managed without change to the bits that make it up and according to relevant policies and standards.

Authenticity: The quality of being genuine and free from tampering and is typically inferred from internal and external evidence, including its physical characteristics, structure, content, and context.[50] Trustworthiness.

Back coat: Layer added to some magnetic tape to help support the magnetic recording layer. The back coat reduces tape friction, dissipates static charge, and reduces tape distortion.

Binder system: System through which magnetic particles are held by a binder to a substrate layer.

Bit rot: The corruption, loss, or decay of bits, the building blocks of digital files.

Carrier type: Refers to the physical carrier of the AV material. Examples of carrier type include reels and cassettes.

Checksums: Alphanumeric strings that reflect the uniqueness of every digital file.

Curation: The activities that are performed on a digital file throughout its lifecycle, including selection and appraisal, description, ongoing care and management, long-term access, and/or deaccessioning/disposal.

Degradation: The process in which the quality or integrity of an object is destroyed over time.

Delamination: In disc media, the process that causes layers to separate from the support base.

Digital preservation: The active management of digital content over time to ensure ongoing access."[51] It is an integral part of curation (see definition above).

Digitization: The representation of an object, image, sound, moving image, or document by generating a series of numbers that describe a discrete set of its points.

File attendance: Ensuring that there are no missing or unexpectedly present files in a given location.

Fixity: File fixity refers to the property of a digital file being fixed, or unchanged. Fixity checking is the process of verifying that a digital object has not been altered or corrupted.[52]

Governance: In the informational sense, governance is the set of structures, policies, procedures, processes, and controls implemented to manage information at an enterprise level, supporting an organization's immediate and future regulatory, legal, risk, environmental and operational requirements.[53]

50 "Authenticity." Glossary of Archival and Records Terminology. Society of American Archivists. *http://www2.archivists.org/glossary/terms/a/authenticity*

51 "About." Digital Preservation. Library of Congress. *http://www.digitalpreservation.gov/about*

52 Wikipedia. File Fixity. *https://en.wikipedia.org/wiki/File_Fixity*. Accessed December 29, 2016.

53 Wikipedia. Information governance. *https://en.wikipedia.org/wiki/Information_governance*

Ingest: The process by which digital files and their associated metadata (called a Submission Information Package, or SIP) is deposited or submitted into a digital repository.

Latency: In computer networking, latency is the time interval between the request for information, such as a digital file, and the retrieval or display of that file to the user by the system.

Machine transport: Playback equipment.

Mandrel: A cylindrical rod placed through a cylinder and used to rotate it for playback.

Media type: AV materials are classified as audio, video, or film during the cataloging and inventory processes.

Metal evaporated tape process: Process in which magnetic particles are vaporized from a solid and deposited onto a substrate layer.

Migration: Converting from one format to another format considered to be of greater stability.

Obsolescence: The state of being which occurs when an object or practice is no longer wanted or used. Usually occurs when a new technology supersedes the old.

Preservation planning: A process by which the general and specific needs for the care of collections are determined, priorities are established, and resources for implementation are identified.

Refreshing: Copying information content from one storage media to the same storage media.[54]

Reproduction method: Method in which a recorded signal is played back from a physical media object.

Risk management: The systematic control of losses or damages, including the analysis of threats, implementation of measures to minimize such risks, and implementing recovery programs.[55]

RPM: Rotations per minute. Used to indicate recording speed for discs and cylinders.

Sidecar file: A file that is stored next to the AV file in the same directory.

Signal path: The route that an audio signal travels from source to output. This may be within a single device (CD to speaker within a stereo system) or within a workflow (original audio recording to reformatted digital file).

Slipping: Tape pack problem in which either single strands or groups of strands are misaligned and migrate to rest against the edge of the flange. May cause edge damage to the tape or film.

Splice: When two ends of a tape or film are joined together using specially formulated splicing tape.

Sticky shed syndrome: A condition resulting from the deterioration of the binder in magnetic tape that results in gummy residues on tape heads during playback.[56]

Storage architecture: The computing and network infrastructure required to store digital files.

Storage capacity: The amount of data a storage device can hold, often measured in gigabytes (GB), terabytes (TB), and petabytes (PB).

Storage media: Devices on which data is stored. These include computer hard disks, optical disk drives, USB drives and other external hard drives, DVDs, and magnetic data storage tapes.

Stylus: A hard point following a groove in a phonograph record and transmitting the recorded sound for reproduction.

Substrate: The backing film needed to support the magnetic recording layer of a magnetic tape.

Tails out: A method for winding tape onto a reel where the end of the tape is on the outside.

54 Digital Preservation Coalition Digital Preservation Handbook Glossary https://dpconline.org/handbook/glossary#R

55 "Risk Management." Glossary of Archival and Records Terminology. Society of American Archivists. *https://www2.archivists.org/glossary/terms/r/risk-management*

56 "Sticky Shed Syndrome." Glossary of Archival and Records Terminology. Society of American Archivists. *https://www2.archivists.org/glossary/terms/s/sticky-shed-syndrome*

NEDCC

NORTHEAST DOCUMENT CONSERVATION CENTER