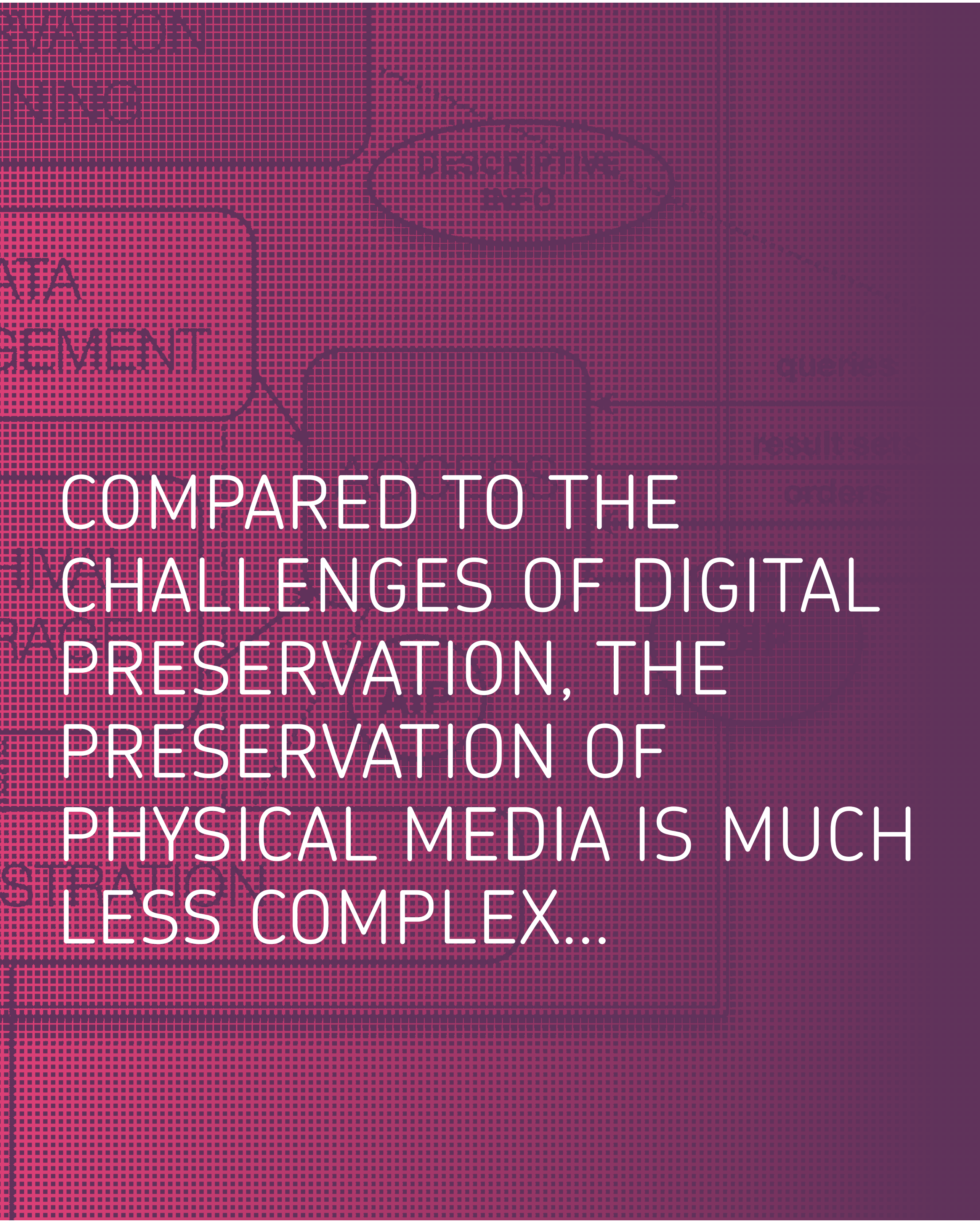
3

KARA VAN MALSSEN

Kara van Malssen werkt als Senior Consultant bij AudioVisual Preservation Solutions in New York. Daarvoor was zij hoofdresearcher en projectmanager van het team dat de preserverings- en distributietechnologie ontwikkelt voor het American Archive van de publieke omroeporganisaties in de Verenigde Staten. Ook heeft Van Malssen gewerkt als Senior Research Fellow aan de Universiteit van New York. Hier was zij betrokken bij het project Preserving Digital Public Television, dat onderdeel vormde van het nationale conserveringssprogramma (NDIIPP) van de Library of Congress. Kara van Malssen treedt regelmatig op als consultant voor organisaties die bezig zijn met het inrichten van digitale repositories en de implementatie van preserverings-metadata, zoals het Museum of Modern Art. Van Malssen is co-chair van de International Outreach Committee van de Association of Moving Image Archivists (AMIA). Zij organiseert, coördineert en draagt regelmatig bij aan internationale workshops op het gebied van opslag, behoud en beschikbaarstelling van digitale audiovisuele content. Van Malssen behaalde een MA in Moving Image Archiving and Preservation aan de NYU.



COMPARED TO THE CHALLENGES OF DIGITAL PRESERVATION, THE PRESERVATION OF PHYSICAL MEDIA IS MUCH LESS COMPLEX...

PLANNING BEYOND DIGITIZATION: DIGITAL PRESERVATION OF AUDIOVISUAL COLLECTIONS[1]

Kara Van Malssen

When we discuss the preservation of digital audiovisual media, we are talking about content that originated in one of two ways:

- The content was born-digital
- The content was digitized from an analog or physical source

Cultural heritage organizations have primarily been concerned with the digitization of analog collections for the past several years. This is understandable, given that millions of hours of physical video and audio materials will be impossible to preserve or access without this important step. However, born-digital collections are increasing becoming part of cultural heritage collections, causing archives' necessarily focus to their attention on issues of collecting, retaining, and preserving file-based materials. These collections will soon become the majority of the archive's holdings. According to the 2010 report, The Digital Universe Decade, "Between now and 2020, the amount of digital information created and replicated in the world will grow to an almost inconceivable 35 trillion gigabytes as all major forms of media – voice, TV, radio, print – complete the journey from analog to digital."[2] A large amount of that data will be audiovisual media, which are some of the largest digital objects out there.

Regardless of whether the content originated on a physical or analog source, or if it is natively file-based, the long-term preservation approaches will be the same. The task of managing large amounts of digital data introduces new

[1] Dit artikel is een bewerking van een presentatie die op 17 november 2010 werd gegeven tijdens het AVA_Net najaarscongres.

[2] John Gantz and David Reinsel, "The Digital Universe Decade – Are You Ready?" IDC, May 2010. Accessed 9 March 2011 from http://www.emc.com/collateral/demos/microsites/idc-digital-universe/iview.htm

challenges for audiovisual archives. Preserving digital audiovisual materials requires new approaches, workflows, tools, resources, and skill sets.

Fortunately, the audiovisual archives community can build on the best practices and standards developed in related fields, including information technology, digital libraries and digital preservation. This report discusses some of those important foundations and describes how they can be interpreted in an audiovisual context. Finally, it offers some strategies for preserving digital audiovisual media, applying the framework offered by relevant standards.

CHALLENGES

Think back to static media, such as a photograph, sculpture, or document. These objects can be viewed with the naked eye – they require no intermediary viewing device. They can be utilized without additional components. And they also have a very long useful life if stored properly. Even motion picture film fits this description to a certain extent.

Magnetic media introduced a new set of obstacles to preservation. These objects require machines in order to be viewed and utilized. The media themselves are fragile, subject to easy damage from temperature and humidity, mishandling, and especially machine malfunction. Frequent industry changes result in format obsolescence, putting the content at risk.

Digital media compound the issues introduced by magnetic media. These media are not tangible – they are simply electronic information composed of bits and bytes, 0s and 1s. They have even more dependencies than the magnetic media did. Whereas magnetic media required a single machine to be played back, digital files require combinations of software and hardware in order to be stored, retrieved, and rendered. The easy corruptibility of digital bits requires ongoing strategies to mitigate loss.

The myriad of file formats, especially for digital video, makes the maintenance of required playback environments difficult. In audio preservation, uncompressed Broadcast Wave Format has emerged as a clear standard for digitization of legacy formats and even the creation of new archival content.[3] Unfortunately, things are not as simple for video. There is no standard container format or codec for digitization. Compounding the problem is the

[3] Recommended by the International Association of Sound and Audiovisual Archivists, TC-04: "Guidelines on the Production and Preservation of Digital Audio Objects" (http://www.iasa-web.org/tc04/audio-preservation) as well as "Sound Directions: Best Practices for Audio Preservation" (http://www.dlib.indiana.edu/projects/sounddirections/papersPresent/index.shtml), amongst others.

range born-digital formats being created today by hard drive cameras. As these files start coming into archives, a wide variety of file formats, codecs, and data rates must be managed.

Ideally, video preservation would also standardize around an uncompressed format. The main reason that this has not happened is that uncompressed video files can be very large, and therefore expensive to store, due to the high data rate (between 216 megabytes per second for SD video, to over 1 Gigabyte per second for HD video). A single hour of uncompressed HD video can be 450 GB. That means if an archive collects or creates 5,000 hours of uncompressed HD video they will need over 2,000 terabytes, or 2.14 petabytes of storage to keep a single copy of each video! Video is compressed in the broadcast and consumer markets to make it easier to store and transmit over the Internet. Archives often make the same choice for their legacy material, and nearly all the file-based material they are accessioning will be born-compressed to some extent. The cost of storage and the speed of bandwidth are improving, though not fast enough for all archival institutions to be able to afford to store uncompressed video.

Many analog formats are in need of migration today, due to deterioration and obsolescence. This problem is quite severe, particularly in light of the fact that there are not enough working analog decks left in the world to transfer all of the videotapes in need of preservation. As Jim Lindner noted in a message to the AMIA listserv, “the small population of decks make it mathematically improbable that a great deal of this work can ever be transferred – there is simply not enough equipment to do it – at any price.” He adds, “ We have lost the chance to save it all – now we must move quickly to identify and save what is critical.”[4]

Some may say this is a gross over-simplification, but compared to the challenges of digital preservation, the preservation of physical media is much less complex. Physical media preservation is primarily about good storage, disaster protection, and a means to locate objects on shelves. Take film as an example: if stored in optimal conditions, it can last for hundreds of years. The same cannot be said for digital media. Much more than just good storage is required.

Finally, one challenge that is not frequently addressed in discussions on digital preservation is the issue of value. Why should people invest in the preservation of digital media? Creating value for the constantly evolving expecta-

[4] Jim Lindner, “End of Quad and One Inch.” Discussion on AMIA-L, 21 May 2009.

tions of stakeholders and users is of critical importance for the sustainability of digital preservation. The authors of the 2009 report, "Sustaining Digital Resources: An On-the-Ground View of Projects Today" note, "Sustaining the value of the resource requires more than just 'keeping the lights on." They add, "As new technologies develop and user expectations shift and grow, a resource risks fading slowly into irrelevance if it does not constantly grow and innovate in ways that continue to benefit its constituents."[5] Keeping up with those changing user expectations is an enormous challenge. Recent literature, however, such as Clay Shirky's Cognitive Surplus (2010) or Yochai Benkler's The Wealth of Networks (2007), offer a starting point for understanding today's networked users, their needs and behaviors.

RISK FACTORS

Digital media face a number of unique risks. Many of these are common to all digital file types, from video to text to databases. These risks are described extensively in digital preservation literature. A good reference is 2005 report, "Requirements for Digital Preservation Systems: A Bottom Up Approach," in which the authors detail 13 threats that digital objects face, summarized as follows:[6]

- Storage media and hardware failure: This commonly results in what is known as "bit rot," which occurs when digital bits flip from 1 to 0 or vice versa, causing potentially catastrophic effects to the media.
- Software failure: Viruses and other problems may render files unreadable.
- Communication errors: File transfer is quite often the time that data corruption takes place.
- Failure of networked services: In a networked environment, high dependence on resources managed elsewhere (i.e. links to external URLS) can cause disruption and failure of resource execution.
- Media and hardware obsolescence: As technology matures and storage density increases, current hardware will soon become obsolete (think of floppy disks and zip drives). Computer processers, specific to many software applications, also become obsolete, and can contribute to a resource becoming unreadable.

[5] Nancy L. Maron, K. Kirby Smith, and Matthew Loy, "Sustaining Digital Resources: An On-the-Ground View of Projects Today." Ithaka S+R, July 2009, 11. Accessed 10 March 2011 http://www.ithaka.org/ithaka-s-r/research/ithaka-case-studies-in-sustainability/report/SCA_Ithaka_SustainingDigitalResources_Report.pdfa

[6] David S. Rosenthal et al, "Requirements for Digital Preservation Systems: A Bottom-Up Approach." D-Lib Magazine, November 2005. Accessed 10 March 2011 http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html

- Software obsolescence: Most software is not maintained in perpetuity. The high dependence on software in order to playback audiovisual files (sometimes one video file depends on a multiple software applications to be read) means that unsupported software, even as a result of software upgrades, can result in playback problems.
- Operator error: A most common anecdotally reported cause of problems in digital environments is human error.
- Natural disaster: As with analog media, digital content is at high risk during disasters such as floods, fire, and earthquakes.
- External or internal attack: Malicious attack by either internal or external forces in a digital environment.
- Economic failure and organizational failure: Often overlooked, this is perhaps the biggest threat of all. If an organization chooses to no longer support the digital preservation environment – either due to bankruptcy, change of mission, or simply a lack of funds – the digital resources risk disappearing.

Another important risk, not identified by the authors of the aforementioned report, is a lack of metadata. Think of how hard it is to organize and keep track of the files on your own computer. Now imagine sharing that computer with a large number of people, all of whom are adding files, and who need to access files for different purposes using different search criteria. Without an agreed upon organizational structure, file and folder naming conventions, and descriptive information users can search upon, retrieving what one is looking for can become an exercise in extreme frustration. In a digital preservation environment, especially where thousands of audiovisual objects are being managed, a lack of metadata is an enormous threat to long-term accessibility of the files. As the authors of Descriptive Metadata for Television note, "If a piece of program material is not correctly placed and identified on a digital system, it might as well not be there – no one will be able to find it or even know it exists."[7]

The fact that many of the threats to digital longevity may elevate the risk of another (i.e. human operator error is more likely under the stress of an internal attack) underscores the need for a managed preservation environment.

[7] Cox, Mulder, Tadic. Descriptive Metadata for Television (Focal Press, 2006), 63.

PRESERVATION REQUIREMENTS

While the threats to sustainable digital preservation of audiovisual materials are numerous, a community of practice in the archives, library, museum, science, commercial, and government sectors has emerged that can offer guidance to those managing digital collections over the long-term. Basic principles of digital preservation can be applied and expanded to meet the needs of complex audiovisual resources.

If we examine the threats identified above, and keep in mind the needs of the user communities for which the content is being preserved, three overarching preservation requirements emerge:

- **Bit Preservation:** In order for digital bits to remain uncorrupted over time, data must backed up on multiple (ideally different) storage media, geographically distributed, periodically "audited" or checked for errors, and protected against security breeches or natural disaster. It is not sufficient to store digital files on one server or hard disk, even one employing RAID[8] technologies. These redundancy methods only protect against a few types of failures, which doesn't include bit rot (corruption), disasters, human error, or viruses. A digital preservation repository will use strategies to protect against common digital threats, actively monitor all copies, and perform restoration in the case of corruption or data loss of the primary copy.

- **Content Accessibility:** This requirement is in place to ensure that video, audio, and ancillary files can be found, retrieved, played back, and delivered to users. It isn't enough to simply keep the files and make sure all the bits are still intact when needed. A managed environment is one that guarantees that files can be identified and located in the system. Good digital file naming conventions are one step in this direction, and relating file names and file locations to a database is another. Descriptive, technical, and structural metadata about the digital object must be collected and/or created, maintained, and updated. The comprehensive documentation and metadata creation/collection practice is one of the largest challenges in digital preservation, but is critical in order for the work to be found and understood over the long-term.

- **Ongoing Management:** Over the past 20 years or so, many digitization projects have been funded and completed at cultural heritage institutions.

[8] RAID = Redundant Array of Independent Disks. It is a term to describe data storage solutions that divide and replicate data over multiple hard disk drives. While it helps protect data from some errors, RAID is not considered a backup solution.

> Unfortunately, once the initial digitization money dries up, so does the institutional commitment for the project. Many institutions have been left with heaps of digital information that is no longer accessible, because the ongoing digital preservation activities were not supported. Due to one or more of the risks outlined above, the digital content is lost. The fact is: there is no starting and stopping point to digital preservation. It is an ongoing process that requires ongoing support of the repository's management. In a larger institutional context, this requirement implies that the institution that houses digital collections is committed to supporting the work of its repository. This includes the sufficient staffing and funding for the repository to perform its functions, as well as instituting and upholding organization-wide practices that enable cost-effective digital preservation.

These three requirements can be found as repeated, overarching themes in the standard literature on digital preservation. One of the most important of these resources is the ISO standard Reference Model for an Open Archival Information System (**OAIS**).[9] OAIS provides a high-level model of the functions, processes, responsibilities, and information required to implement a digital preservation repository. It also defines mandatory responsibilities expected of a digital repository, including negotiating and accepting appropriate information from creators, obtaining sufficient control of the information to meet long-term preservation, and following documented policies and practices to ensure preservation of information against reasonable contingencies.

Repository audit and certification criteria are also important resources. These include the Trusted Repositories Audit and Certification: Criteria and Checklist (**TRAC**)[10] by the National Archives and Records Administration (NARA) and the Center for Research Libraries (CRL), the Nestor Catalogue of Criteria for Trusted Digital Repositories,[11] and the DRAMBORA[12] risk assessment toolkit jointly created by Digital Preservation Europe and the Digital Curation Centre in the UK. Work is currently underway to create an ISO standard for audit and certification based on these criteria. All address in detail the requirements of digital conservation – bit preservation, content accessibility, and ongoing management – given above.

The Preservation Metadata Implementation Strategies (**PREMIS**)[13] metadata standard is also a useful reference for identifying the important as-

[9] http://public.ccsds.org/publications/archive/650x0b1.PDF
[10] http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf
[11] http://www.dcc.ac.uk/resources/tools-and-applications/nestor
[12] http://www.repositoryaudit.eu/
[13] http://www.loc.gov/standards/premis/

pects of digital files and digital preservation environments that should be documented.

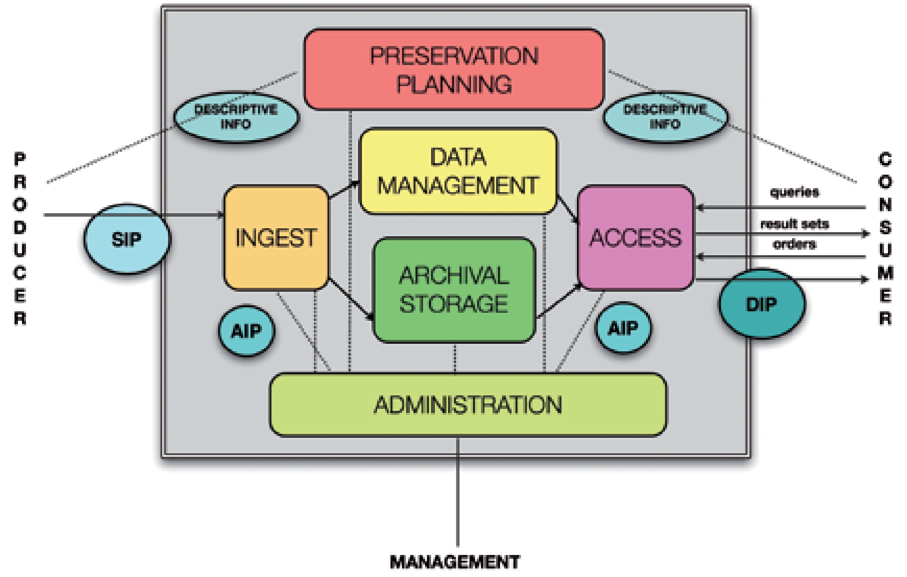
Each of these resources emphasizes the central role that authenticity plays in the function of a digital preservation service (as opposed to another type of digital service, which does not have a preservation mission). Maintaining the authenticity of digital objects, whether they are born-digital or digitized from an analog source, remains a fundamental task of the digital archive. TRAC lists as one of its criteria, "Repository enables the dissemination of authentic copies of the original or objects traceable to originals."[14] PREMIS further elaborates, "Authentication, or the demonstration of authenticity… includes both technical and procedural aspects. Technical approaches may include the maintenance of detailed documentation of digital provenance (the history of the object), the preservation of a version of the object, that is, bit-wise, identical to the content as submitted."[15]

For born-digital video files, ensuring authenticity might mean guaranteeing that the native resolution, colorspace, or data rate is preserved and not compromised during a migration event. For analog materials that are being digitized, it might mean choosing an encoding format that best represents the technical specifications of the original. For dissemination purposes, it might mean communicating important provenance information to users: What was the original format? What type of camera was used to record the footage? How does an available proxy differ from the original audio file? If the user, perhaps a broadcast producer, is viewing a low resolution proxy with visible artifacts, it might be important for them to know whether those artifacts were inherent to the original footage, or simply a result of the quality of the streaming copy.

OAIS, TRAC and PREMIS offer helpful starting points from the general digital preservation community, which can be built upon and refined for the needs of audiovisual collections. They will be referenced throughout the remainder of this paper.

[14] RLG-National Archives and Records Administration Digital Repository Certification Task Force, "Trustworthy Repositories Audit and Certification: Criteria and Checklist." Center for Research Libraries/OCLC, 2007, 41. Accessed 10 March 2011 http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

[15] PREMIS Editorial Committee, "PREMIS Data Dictionary for Preservation Metadata" version 2.1, January 2011, 210. Accessed 10 March 2011 http://www.loc.gov/standards/premis/v2/premis-2-1.pdf



The OAIS functional model

INTRODUCTION TO OAIS

OAIS covers the fundamental terminology, concepts, and framework for the long-term store and access of digital objects and their associated metadata. It is general enough to be applicable to all digital collections, regardless of the type of content being preserved. This introduction to OAIS is offered to give the reader a reference for the strategies for digital audiovisual preservation outlined in the following section, which are framed using OAIS terminology.

The OAIS reference model describes three areas that collectively make up a repository's operation: the external environment (producer, consumer, and management), the functional components of the repository itself, and the information packages being preserved and disseminated. Each part plays equally important roles in the long-term conservation and access of digital information.

EXTERNAL ENVIRONMENT

The environment outside an OAIS repository has a critical impact on the internal functions, policies, practices, and methods for conservation and dissemination of content. The Producer community includes anyone outside the repository responsible for the creation of content. This group could include a producer in the traditional broadcast sense, an artist, scholar, or even curators responsible for acquisition within the institution, though outside the repository. Management is the overseer, funder, and strategic planner of a conservation repository. The Consumer community includes any user of content in the repository, from administrators and curators, to educators, creators, and the general public (note that in many cases the Producer community is the same group as the Consumer community).

FUNCTIONAL ENTITIES

OAIS defines six internal functions of a digital repository. Collectively, these functions are in place to ensure that digital information is conserved and made accessible over the long-term. They are groups of processes that must be present and working together in order to systematically fulfill the repository's digital conservation mission. It is likely that multiple departments in an organization could be involved in one or more functional requirement, and that one staff member can fulfill more than one function. Keep in mind, however, that these functions are specific to the repository's unique role of digital preservation. Although the functions in digital preservation involve interaction with the wider organizational environment, they should be considered distinct from other existing workflows. The six functions of an OAIS repository work with the external environment to safeguard digital information packages.

- **Ingest:** Brings the content into the repository.
- **Archival Storage:** Manages storage, periodically checks files for errors, maintains backups, facilitates repair of corrupted files.
- **Data Management:** Administers the database that contains information about the repository's holdings.
- **Preservation Planning:** Is responsible for planning, reviewing, and updating the repository's preservation strategy.
- **Access:** Facilitates requests to archival storage and data management, generates Dissemination Information Packages, and delivers the information in the appropriate format to the users.
- **Administration:** Oversees the operation of the entire system.

INFORMATION MODEL

The information packages that are acquired, conserved and disseminated by a repository include a digital object (a video interview with an historic figure, for example) along with associated metadata (time and place of interview, events and personalities mentioned in the interview, etc). The components of the information package help to ensure that the object can be managed, located, authenticated, and interpreted. There are three versions of the information package that are transformed through preservation process: the Submission Information Package (SIP), the Archival Information Package (AIP), and the Dissemination Information Package (DIP).

- **Submission Information Package:** The SIP is the package acquired from the submitter (the contributor and their agents), contains the essence (video and audio) files and the minimum metadata required by the repository.

- **Archival Information Package:** The AIP is the complete archival object, and includes descriptive, technical, administrative and/or preservation metadata, much of which is added to the SIP by the repository. This metadata helps the repository better manage and provide access to the content. The AIP may also include new digital audiovisual files created by the repository, including derivatives such as mezzanine (working copies) and proxies (access copies), and potentially transcoded preservation master files, depending on the repository's policies.

- **Dissemination Information Package:** The DIP varies with each user and use, but it is generally a limited version of an AIP, since users of a repository don't need or want to know every detail about content maintained in an AIP. Thus, the DIP is less a replica of the AIP than simply a portion of it. DIPs will vary for each distribution platform that the repository is delivering to.

DIGITAL PRESERVATION STRATEGIES FOR AV COLLECTIONS

In this section, the OAIS reference model is used as a starting point to identify points in the preservation workflow of audiovisual archives where strategies might be used to help achieve the three requirements of digital preservation: bit preservation, content accessibility, and ongoing management. For each OAIS functional entity (i.e. Ingest, Data Management, etc) a handful of strategies are offered specifically for audiovisual archives. Though this is certainly not an exhaustive list of approaches, these strategies are general enough to be applicable to nearly all organizations faced with long-term management of digital video and sound collections. The strategies discussed here are informed by my experiences planning developing OAIS-based audiovisual preservation repositories at New York University, the American Archive of public broadcasting, and the Museum of Modern Art's Digital Repository for Museum Collections.

As described above, in contrast to many other file types, digital audiovisual files are large, complex, and potentially very costly to store, manage, and disseminate. Audiovisual archives are faced with new and daunting challenges as they move into the digital realm, both technical and economic. As born-digital submissions increase, selection criteria, submission requirements, workflows for ingest, in-house standards, and access protocols become increasingly important. The goal of any preservation strategy for audiovisual collections should be threefold: ensure efficiency, reduce costs, and maintain authenticity.

INGEST

Ingest is the point of entry into the repository. The actions taken here have tremendous implications for all other functional entities. As the ingest entity is responsible for bringing content into a digital repository, it must be sure that submissions can be managed once the ingest process complete. Given the daunting number of file formats in use in audiovisual production environments today, a repository cannot be expect to manage all of them over time long-term. Additionally, as creators know their content best, a repository cannot efficiently nor cost-effectively preserve and provide access to content that is submitted without any descriptive metadata. Thus, one of the most important tasks of a digital repository is to create Submission Information Package (SIP) requirements, and enforce these at ingest.

The creation of SIP requirements is a critical step in a digital preservation repository's development. They reflect the repository's internal capabilities to provide preservation services. To ensure smooth ingest, files must be easily managed in a tested workflow, given the tools and expertise available. Submission guidelines will help ensure that content is uniform, bottlenecks are reduced, files are interoperable, and processing costs are kept within available budgets.

SIP requirements for AV collections should specify the file wrapper formats and codecs that are accepted. This involves an examination of both in-house toolsets, as well as the repository's level of commitment to preserve a particular format. If the repository has a toolset that works better with files wrapped in the MOV file wrapper format (as opposed to MXF or AVI), it should require that submissions be wrapped as MOV. If the repository can manage all of these file wrappers, this might not be an issue. Video codecs,[16] however, are numerous and complex, some of which may not easily fit a workflow, or be conservable after a few years. A repository should identify the common codecs that its producer communities create, and decide what can be accepted. Will the repository accept all codecs? Or only a limited, identified set? Can it accept popular but proprietary (and continually evolving) editing formats like Apple's ProRes[17] or Avid's DNx.[18] Or will it be limited to more open and standard formats like DV, MPEG-2, or even uncompressed? Codecs that are supported

[16] An extensive list of video codecs is available from https://secure.wikimedia.org/wikipedia/en/wiki/Video_codecs#Video_codecs

[17] https://secure.wikimedia.org/wikipedia/en/wiki/ProRes_422

[18] https://secure.wikimedia.org/wikipedia/en/wiki/DNxHD_codec

by toolsets with a wide community of users, such as FFMPEG's libavcodec[19] library, are much more likely to be sustainable than commercial formats.[20]

Repositories must be equally concerned about metadata. If content is submitted without minimal descriptive information, the repository could potentially spend unnecessary time and resources (and could also risk infringing the authenticity of the object) by having to research basic information like title, creator, and description, or figuring out what the rights status is. Efforts like these can prevent smooth automated ingest workflow. They may prevent the repository from investing its resources in enhancing metadata records and improving content accessibility, simply because so much time is spent trying to create basic records, and reducing backlog. Even requiring a very limited number of fields to be submitted in a specified format can help improve the ingest process tremendously. Guidelines and tools may need to be provided to submitters to facilitate this process.

By following the repository's SIP requirement policy, the ingest entity helps ensure that submitted content can be managed effectively by the other functional entities over time.

ARCHIVAL STORAGE

The Archival Storage Entity manages storage and backups, ensures the integrity of digital files, and maintains security of the system. As no digital storage solution is fail-proof, it is always necessary to create multiple copies of files, and store them in separate (ideally, geographically separate) locations. There are various options for storage – the choice of solution should balance volume with the frequency of access required and total cost (including power and climate control required). Preservation master files may be accessed very rarely, so if the repository has a large volume of digital content perhaps one copy is stored in a data tape robot and the backups are on data tapes on shelves, which don't require any ongoing power supply. Proxy files, those created specifically for access purposes, may be stored online so they can be quickly retrieved, in a local NAS[21], file server, or even in a cloud storage service like Amazon S3.[22]

Archival Storage must be prepared to monitor the files and periodically check them for corruption. A standard method for doing this is to document

[19] http://www.ffmpeg.org/

[20] The Library of Congress offers a set of sustainability factors for file formats that can help a repository weigh the longevity of a given format. See http://www.digitalpreservation.gov/formats/sustain/sustain.shtml

[21] NAS = Networked Attached Storage. See https://secure.wikimedia.org/wikipedia/en/wiki/Network-attached_storage for more detail

[22] http://aws.amazon.com/s3/

a checksum for each file (ideally, these checksums were generated before the file was submitted, so that they can be checked upon ingest), then periodically "audit" files using the same checksum in order to detect any changes. A checksum is the mathematical signature for a media file, a receipt that can prove all of its bits are in order. If at any time bits of the media file are lost or corrupted, the new checksum will no longer match the original, so the file needs replacement by one of the exact copies stored elsewhere. This process will go on for the life of a particular media file.

The checksum and repair process is common to all digital preservation environments. However, this method is imperfect for audiovisual content. If a checksum change is detected for a particular video file, the operator has no way of determining exactly where the damage is, and is forced to replace the entire file, a time consuming and resource intensive process, especially if file errors are detected for multiple files. Fortunately, new technologies and practices are emerging which will help make the integrity management process more efficient for audiovisual files. As the UK Avatar-M project demonstrates, by "chunking" or segmenting files for storage, sophisticated archiving processes can be enabled, which, "recognizes that not all parts of an AV asset are 'equal' when it comes to preservation."[23] Using this method, checksums can be applied to each chunk. Thus when a repair event takes place, only the corrupted chunk must be replaced, rather than the entire file. As tools are developed to support chunking of AV files, this method holds a lot of promise, especially for large archives.

The archival storage entity is responsible for storing the entire Archival Information Package (AIP), which, as previously mentioned, will likely include more information that was in the SIP. For example, the repository will need to add its own administrative metadata, it may extract technical metadata from files, and/or may enhance descriptive metadata records with more detail than was provided by the submitter. It also may have created additional audiovisual files, mezzanine and proxy files, that can be used to facilitate access, so that the important (and bulky) preservation master can be left alone. Much like the SIP requirements, an AIP specification, that includes file and folder naming conventions, metadata requirements, metadata standards, and file types (i.e. preservation, mezzanine, proxy) expected, can both help automate workflows as well as ensure consistent data indexing by the data management entity, described belpw. An administrative file wrapper, such as METS,[24]

[23] Matthew Addis, Richard Lowe, Lee Middleton, "A New Approach to Digital Audiovisual Archiving," Presented at NAB 2009, 4. Accessed 21 February 2011 http://www.it-innovation.soton.ac.uk/projects/avatar-m/addisnab2009.pdf

[24] http://www.loc.gov/standards/mets/

or directory structure, such as BagIt,[25] will help consistently package AIPs. Different AIP classes may be detailed for different file formats, or content types, to enable consistent processing of data.

DATA MANAGEMENT

This entity administers the database about the repository's holdings. It is responsible for making sure content and information about that content is accessible to administrators, curators, and diverse user communities. All other functional entities depend on the information managed by this entity. Access to the archive will only be as good as the metadata that is collected, created, and made available to dissemination platforms. Preservation planning will greatly benefit from technical metadata that identifies formats, file size, and data rates of files. Administration will need deposit and usage analytics to help with budget and project planning.

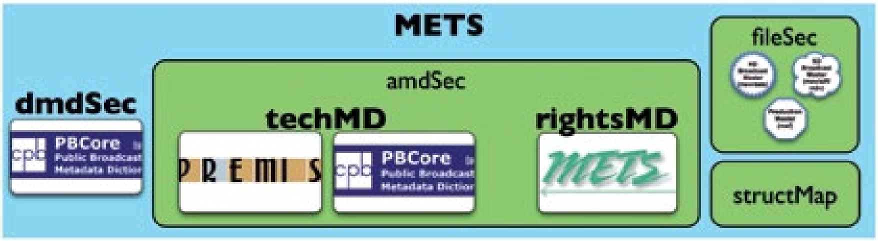
Good metadata should be maintained throughout the lifecycle of the digital object. This means requiring essential descriptive and rights metadata to be created and submitted along with the audio or video essence as part of the SIP requirements. Additional technical and preservation metadata will need to be generated and stored by the repository. This might include extracting technical metadata from files using free tools like MediaInfo,[26] and storing that information in a standardized format. It also includes developing an approach for documenting fixity checks (i.e. running checksums to check for corruption) and migration events. The data management entity collects and makes this information available to the other entities.

The data management entity must be prepared for metadata to change over time. For example, rights and permissions status may change with the passing of a new law, or a submitter deciding that they want to blacklist another organization from using their content. As access to audio and video materials is dictated by complex intellectual property laws, it is important that the repository be prepared to respond to such changes appropriately.

New descriptive metadata might be continuously added. Although, catalogers and loggers are the authoritative creators of descriptive information, but there is only so much time and only so many resources available to allow humans to watch and describe video, or listen to audio recordings in full. As technology advances, tools for automated audio transcription and video object recognition will improve, and improve the discoverability of these col-

[25] http://www.digitalpreservation.gov/partners/resources/tools/index.html#b

[26] http://mediainfo.sourceforge.net/en



A diagram of an example AIP from the Preserving Digital Public Television Project. METS is the administrative wrapper for the AIP. PBCore is used for descriptive (dmdDec) and technical (tmdMD) metadata. PREMIS is used to capture additional technical metadata. Rights are documented using the METSRights standard. Files include an HD broadcast master a SD broadcast master, and a production master, all requirements for this particular AIP class. From "PBCore, METS, PREMIS, MODS, METSRights... oh my!" Presented at the Association of Moving Image Archivist Conference, November 2008.*

* http://www.slideshare.net/kvanmalssen/pbcore-mets-premis-mods-metsrightsoh-my

lections. This machine-generated information can greatly improve access to under-cataloged or un-transcribed material.

Another source of metadata might be the users themselves. As the web becomes increasingly participatory, it is of value to audiovisual archives and their users to enable the creation of user-generated metadata, such as tags, comments, description, and ratings.

Data management must ensure that all of this information is collected consistently and normalized to an internal metadata structure or schema so that it can be managed, indexed, and queried consistently. This metadata structure is an important part of the overall Archival Information Package.

As mentioned previously, the AIP is the full package of content including essence files (video, audio, images) and associated metadata for a given content item. In addition to consistent AIP packaging, it is important that the components of the AIP also are consistent. The application of metadata standards, such as PBCore[27] or EBU Core[28] for descriptive and technical metadata, PREMIS for preservation metadata, and ODRL[29] for rights metadata, will enable the repository to capture detailed, extensible metadata consistently

[27] http://pbcore.org/
[28] http://tech.ebu.ch/lang/en/MetadataEbuCore
[29] http://odrl.net/

throughout the content lifecycle. The creation of in-house profiles of these standards, defining which fields are required and which are optional and identifying controlled vocabularies that fit the repository's content requirements, will further refine the AIP rules for the context.

ACCESS

The information delivered to users is known as a Dissemination Information Package (DIP). DIPs are typically a sub-set of the canonical Archival Information Package held by the repository, and vary between users and access platforms. The needs and expectations of users today are quite varied. Some users, such as producers, may want to see high-resolution video files and complete transcripts. Others, such as educators, may specifically require low-resolution video files (due to low bandwidth availability in classrooms) and curriculum guides to accompany the archival content. Meeting the needs of all users is an enormous challenge. Utilizing existing infrastructure, services, and products for dissemination of audiovisual content can relieve the archive of a few expensive infrastructure requirements, while simultaneously expanding access to diverse user communities.

Providing access to archival digital audio and video can be an enormous challenge. As archival video files can be particularly large, it can be difficult to move them over file networks or deliver via the web. Audiovisual archives necessarily create proxies – lower resolution access copies – that can more economically and efficiently be streamed to the web. However, even delivering large numbers of even these smaller videos can become burdensome. Like video distributors, audiovisual archives, especially those anticipating high numbers (over 100) of concurrent users, may need to utilize a content delivery network (CDN) to help facilitate access. Per Wikipedia, "A content delivery network or content distribution network (CDN) is a system of computers containing copies of data, placed at various points in a network so as to maximize bandwidth for access to the data from clients throughout the network. A client accesses a copy of the data near to the client, as opposed to all clients accessing the same central server, so as to avoid bottlenecks near that server."[30] There are a number of CDN solutions to consider exploring.

Rather than attempt to develop platforms to reach all users, or assume that a single access point will satisfy diverse user needs, it may be advantageous to develop partnerships for reaching various user groups. As the majority of users are regularly using sites like YouTube and Facebook to find and

[30] "Content delivery network." Wikipedia. Accessed 10 March 2011 https://secure.wikimedia.org/wikipedia/en/wiki/Content_delivery_network

watch video, adding content to these sites can be a great way to effectively reach wide audiences and increase awareness of an archive's collections. For European organizations, participation in Europeana is a powerful way to contribute to the aggregation of all types of content from cultural heritage institutions, providing an excellent one-stop-shop for researchers, educators, and the general public.

Specific user communities can be targeted through strategic partnerships. For example, if all the primary schools in your region have already subscribed to an educational video distribution service, it may be worth exploring how your archive's content can be made available through that service, rather than try to compete with them. Likewise, the producers in your region may be accustomed to visiting specific stock footage providers to find content they need for new productions. By adding your archive's content to the pool of material available through these providers, you may be able to increase licensing revenue, and offset preservation costs (pending rights clearance, of course).

By offloading a portion of access to those that already have built the infrastructure to support and enhance delivery for users, the archive can focus it's attention on developing additional services and adding value. Mobile applications (iPhone, iPad, Android), APIs, and topical or thematic websites are all areas that archival institutions are starting to investing in, finding ways of broadening the outreach of audiovisual content.

PRESERVATION PLANNING

The preservation planning entity ensures that the content will be accessible today, 10 years from today, and beyond. It is responsible for monitoring the landscape, remaining aware of technological advances that may affect an archive's digital collections, and making decisions up front that will enable the repository to be effectively and economically prepared to keep pace with changes as needed, without compromising the integrity and authenticity of the digital object.

An obvious example of a task that the preservation planning entity would be responsible for is the preparation and execution of format migration. As new formats gain popularity and others lose software support, a repository might decide that it is in their best interest to migrate, or transcode, files in their collection to a new format. This action may not be necessary for archival files that are in uncompressed formats, but probably will be periodically required of any access copies. Thinking back over the last 15 years, a number of once popular video streaming formats have come and gone: Real Media, Windows Media, and now Flash. The online video world is currently experiencing a battle over whether H.264 (supported by Apple and Adobe) or WebM (supported by Google and Wikipedia) will emerge as the standard video codec for delivery over the web via HTML 5 (only two years ago Ogg Theora was considered the favorite; now that format has all but disappeared from the debate).

There are certainly more such battles to come, especially in the broadcast environment, where manufacturers of high definition cameras are competing for format dominance. Archives that accept born-digital files in these formats may choose to "normalize" them to an in-house standard upon ingest, or will need to plan for eventual migration to a standard file format if the original source format loses support in the market. Good preservation planning will give the repository an approach for dealing with these questions.

All that preservation and technical metadata created, captured, and stored by the data management entity will greatly facilitate preservation planning. Continuing with the migration example, having the ability to quickly identify how many of a particular endangered format are housed in the repository, budgets and time estimates can be made for migration plans.

A preservation policy is an important document to that can support preservation decisions. This statement should describe the approach to be taken by the repository for the preservation of ingested objects. It can describe different approaches that will be taken for different submitted formats, whether they can be fully supported (including ongoing future migration) or maintained at a bit-level only. For instance, the repository may determine that its policy for proprietary file formats (such as Apple's ProRes) will be to normalize those at files to a more open preservation format. Conversely, the policy may state that these files will be retained in their original format and evaluated for migration 5 years from ingest.

ADMINISTRATION

Administration oversees the entire system, ensuring continuous, efficient, and reliable preservation. By making sure all other functions of the repository are working together, performing their required roles, the bits will remain stable, and the content accessible. But how can the repository ensure that it is performing the functions effectively? Furthermore, how can a repository demonstrate its trustworthiness to content creators and rights holders who are relying on the digital preservation service?

Understanding all the components of a digital preservation system and measuring how well each of these is performing is yet another challenge. Fortunately, colleagues in other disciplines concerned with digital preservation have developed a few helpful risk assessment tools that digital audiovisual repositories can utilize as well. These include Trustworthy Repositories Audit and Certification: Criteria and Checklist (TRAC) and the Nestory Criteria — both of which provide the foundation for the forthcoming ISO Recommended Practice on Audit and Certification of Trustworthy Repositories.[31] These criteria were developed collectively by implementers of digital repositories who wanted to have a measurable method of demonstrating OAIS-compliance. Whether or not a repository seeks official certification, these criteria provide repository administrators a means to evaluate internal approaches to the three essential aspects of a long-term digital preservation system:

- **Digital Object Management:** This area covers the repository functions, processes, and procedures required to ingest, manage, and provide continuing access to content. For example, does the repository clearly specify, "the information that needs to be associated with digital material at the time of its deposit (i.e., SIP)"?[32]

- **Infrastructure and Security Risk Management:** "These criteria measure the adequacy of the repository's technical infrastructure and its ability to meet object management and security demands of the repository and its digital objects."[33] They include general system infrastructure requirements, appropriate technologies to provide access to users, security and disaster protection.

- **Organizational Infrastructure:** The criteria in this category measure the overall governance and organizational viability, staffing, procedural accountability, financial sustainability, and contracts and licenses. For example, does the repository have "a formal succession plan, contingency plan, and/or escrow arrangements in place in case the repository ceases to operate or the governing or funding institution substantially changes its scope"?[34]

By performing a self-assessment using the criteria outlined in these documents, a repository of digital audiovisual materials should be able to ascertain its own trustworthiness, or identify what areas need improvement.

[31] http://wiki.digitalrepositoryauditandcertification.org/bin/view

[32] TRAC, 22

[33] TRAC, 43

[34] TRAC, 10

CONCLUSION

In conclusion, there is no need to reinvent the wheel. Existing standards like OAIS, TRAC and metadata standards like PREMIS can be repurposed to create AV-specific profiles and solutions. Every institution or consortium will take a different approach to the implementation of these guidelines. The important thing is to remain aware of the solutions that are being developed by the broader digital preservation fields, and contribute to that community of practice.

Digital preservation is still a nascent, emerging field. All types of organizations invested in the long-term preservation of heritage face similar challenges. Partnerships with like-minded organizations are an important way that digital repositories can share challenges and solutions, tools and technologies, and even infrastructure. The PrestoPRIME[35] consortium in Europe is an example of audiovisual archival organizations joining together to collectively address large-scale digital audiovisual preservation problems. The project's new AV Competence Centre will be a key resource for organizations large and small looking to network with others concerned with sustainable audiovisual preservation.

[35] http://www.prestoprime.org/index.en.html