Chapter 6: Preservation Target Formats and Systems

6.1.1 Introduction

6.1.1.1 The following information on the management, long term storage and preservation of digitally encoded audio is based on the premise that there is no ultimate, permanent storage media, nor will there be in the foreseeable future. Instead, those managing digital audio archives must plan to implement preservation management and storage systems which are designed to support processes that go with the inevitable change in format, carrier or other technologies. The rate and direction of technological change is something over which archives have no control and very little influence. The aim and emphasis in digital preservation is to build sustainable systems rather than permanent carriers.

6.1.1.2 The choice of technological storage system is dependent on many factors, of which cost is but one. Though the type of technology selected for preserving a collection may differ according to the specific circumstances of the individual institution and its circumstances, the basic principles outlined here apply to any approach to management and long term storage of digital audio.

6.1.2 Data or Audio Specific Storage

6.1.2.1 To effectively manage and maintain digital audio it is necessary to transform it to a standard data format. Data formats are the file types, such as .wav, BWF, or AIFF, which computer systems recognise. These files, unlike audio specific carriers, technologically define the limits of their own content and are generally encoded in such a way that a loss of data is recognised and remedied by the host system. IASA recommends the use of BWF as defined in Section 2.8 File Formats.

6.1.2.2 Audio specific recording formats which have been available in the past include DAT (Digital Audio Tape) and CD-DA (Compact Disc-Digital Audio). DAT, though once largely used for the remote or field recording of 16 bit, 48 kHz audio is now an obsolete recording system. IASA recommends that any significant content recorded on DAT tape be transferred to a more reliable storage system in accordance with the guidance provided in section 5.5 Reproduction of Digital Magnetic Carriers.

6.1.2.3 The recordable compact disc can be used to record audio in either audio-only (CD-A or CD-DA) or data (CD-ROM) formats. In CD-DA format the encoded digital audio resembles an audio stream and so does not have the advantages of a closed file such as might be recorded on the CD-ROM formatted disc. In the latter though, less data can be stored on the same amount of disc space. IASA does not recommend recording audio in CD-DA form as a preservation target format. There are considerable risks associated with using a recordable CD as a target format in any form and those risks are outlined in Chapter 8 Optical Disks: CD/DVD Recordable. The ever reducing prices and increasing reliability of data management and storage systems make media specific storage approaches, such as CD-R, unnecessary, or at least uneconomic.

6.1.3 Principals of Digital Preservation

6.1.3.1 Digital Mass Storage Systems (DMSS) Principles

6.1.3.2 The following information is based very closely on the practical aspects of Data Protection Strategies from the UNESCO Guidelines for the Preservation of Digital Heritage. It is modified only to reflect the possibility of systems that incorporate non-automated back up, and to reflect the single format concerns of audio digital preservation. The section is included with the kind permission of the author (Webb 2003:16.13).

6.1.4 Practical Aspects of Data Protection Strategies

6.1.4.1 There is a reasonably standard suite of strategies used to manage data in long-term storage. Most are predicated on an assumption that the data carrier itself does not need to be preserved, only the data. The following comprises, in part, those strategies.

6.1.4.2 **Allocation of responsibility:** Someone must be given unambiguous responsibility for managing data storage and protection. This is a technical responsibility requiring a particular set of skills and knowledge as well as management expertise. For all collections, data storage and protection require dedicated resources, an appropriate plan and must be accountable for these strategies, and even very small collections must have access to the necessary expertise and a dedicated person responsible for that task.

6.1.4.3 **Appropriate technical infrastructure to do the job**: Data must be stored and managed with appropriate systems and on an appropriate carrier. There are digital asset management systems or digital object storage systems available that meet the requirements of audio digital preservation programmes, some approaches to which are discussed below. Once requirements have been determined, they should be thoroughly discussed with potential suppliers. Different systems and carriers are suited to different needs and those chosen for preservation programmes must be fit for their purpose.

6.1.4.4 The overall system must have adequate capabilities including:

6.1.4.5 **Sufficient storage capacity:** Storage capacity can be built up over time, but the system must be able to manage the amount of data expected to be stored within its life cycle.

6.1.4.6 As a fundamental capability, the system must be able to **duplicate data as required without loss**, and transfer data to new or 'refreshed' carriers without loss.

6.1.4.7 **Demonstrated reliability** and technical support to deal with problems promptly.

6.1.4.8 **The ability to map file names** into a file-naming scheme suitable for its storage architecture. Storage systems are based around named objects. Different systems use different architectures to organise objects. This may impose constraints on how objects are named within storage; for example, disk systems may impose a hierarchical directory structure on existing file names, different from those that would be used on a tape system. The system must allow, or preferably carry out, a mapping of system-imposed file names and existing identifiers.

6.1.4.9 The ability to **manage redundant storage.** As digital media has a small, but significant failure rate, redundant copies of files at every stage are a necessity, especially the final storage phase.

6.1.4.10 **Error checking.** A level of automated error checking is normal in most computer storage. Because audio and audio-visual materials must be kept for long periods, often with very low human usage, the system must be able to detect changes or loss of data and take appropriate action. At the very least the strategies in place must alert collection managers to potential problems, with sufficient time to allow appropriate action.

6.1.4.11 Technical infrastructure must also include means of **storing metadata and of reliably linking metadata** to stored digital objects. Large operations often find they need to set up digital object management systems that are linked to, but separate from, their digital mass storage system, in order to cope with the range of processes involved, and to allow metadata and work interfaces to be changed without having to change the mass storage.

6.1.5 Philosophy of System Sustainability

6.1.5.1 All technology, whether it be the hardware or software, formats or standards, will eventually change as a result of market forces, performance requirements or other needs or expectations. The task of the audio archivist charged with maintaining digital and digitised audio content is to navigate a way through these technological changes such that the content of their collections are maintained for current and future users in a reliable and authentic form in as cost effective way as can be managed.

6.1.6 Long Term Planning

6.1.6.1 Long term planning for a digital audio archive involves more than just the technical standards for a data storage system. The technical issues must be carefully resolved, but the social and economic aspects of running a digital storage system are vital to ensuring the continued access to the content. Long term planning should consider the following issues.

6.1.6.2 **The sustainability of the raw data:** that is the retention of the byte-stream in its proper and logical order. The data in the storage system must be returned to the system without change or corruption. It is worth noting that computer systems expertise identifies a considerable risk in the maintenance and refreshment of data, and only a well managed and designed approach to IT will ensure adequate results.

6.1.6.3 **Formats and ability to replay**: Digital data is only useful in a sound archive if it can be rendered as audio in the future. The proper choice of file format ensures that the future sound archive can replay the content of the data files, or will be able to acquire the technology to migrate the files to a new format. Not incorporating a lossy compression algorithm in that format allows that future transformation process to occur without altering the original audio content.

6.1.6.4 **Metadata, identification and long term access:** All digital audio files must be identifiable and findable in order for that audio material to be used and the value of the content realised.

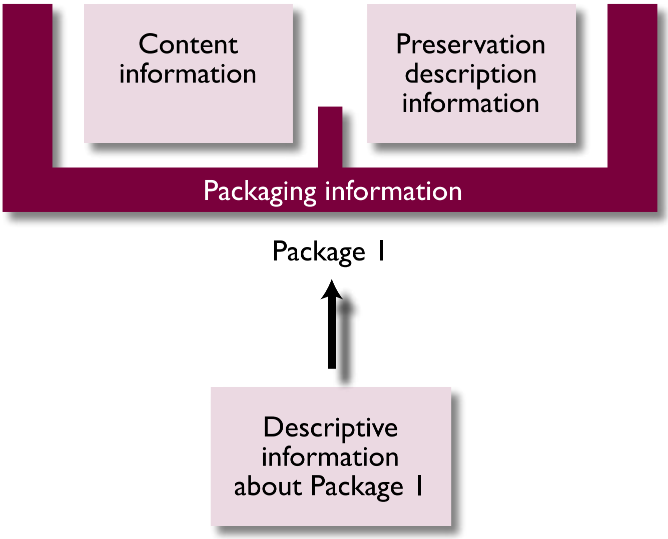
6.1.6.5 **Economics and Sound Archives**: this includes the continued viable existence of the institutions that support the data storage systems and repositories as well as those that own, manage, or gain value from, the digital audio stored therein. The cost of maintaining a digital audio collection is ongoing and their must be a plan and a budget that realistically plans for long term preservation of collections. The cost of curating and managing the audio collections is also ongoing. Digital preservation is as much an economic issue as a technical one. The requirements of ongoing sustainability demand at their base a source of reliable funding, necessary to ensure that the constant, albeit potentially low level, support for the sustainability of the digital content and its supporting repositories, technologies and systems can be maintained for as long as it is required.

6.1.6.6 **Storage, management and preservation alternatives**: Given that the economic and technical environment may well be volatile it is recommended that agreements be established between archives and institutions regarding the storage of data as archives of last resort. This would require some standard agreement about file formats and data organisation as well as social and technical aspects of management of the content.

6.1.6.7 **Tools, Software and long term planning:** Hardware, software and systems are not things in themselves to be preserved, but are merely tools to support the task of preserving the content. The repository software D-Space, for example, does not describe itself as a preservation solution, but only useful in "enabling institutions with a sustainable ability to retain information assets and offer services upon them." (DSpace, Michael J. Bass et al. 2002). The repository software itself is a tool, as are the various components designed to aid in operation, simplify processes, and automate and validate the harvesting of metadata. Long term planning involves being able to change or upgrade systems without endangering the content.

6.1.7 Defining the Digital Object

6.1.7.1 The audio file is only one part of the information that is to be preserved. The Reference Model for an Open Archival Information System (OAIS) identifies four parts to the digital object, described by them as the information package. These are the content information and the preservation description information, which are packaged together with packaging information, and which is discoverable by virtue of the descriptive information.


Content information
Preservation description information
Packaging information
Package 1
Descriptive information about Package 1

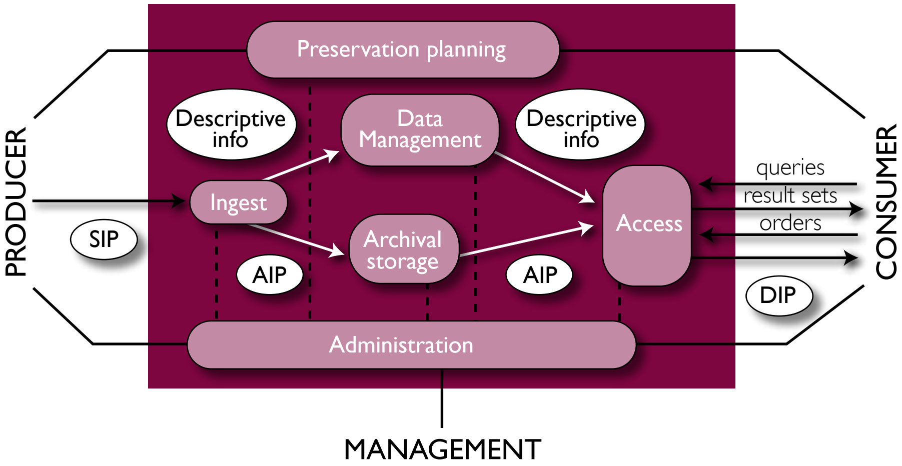
Information Package Concepts and Relationships

6.1.7.2 Though the information may be distributed across the storage system, it is well to remember that the conceptual package is the audio information, the ability to replay that audio, to know its provenance and to describe and find it. There may also be critical relationships between the one audio file and others in the collection, and these relationships are important to the use of the material and so must also be preserved.

6.1.8 The Open Archival Information System (OAIS)

6.1.8.1 The Reference Model for an Open Archival Information System (OAIS) is a widely adopted conceptual model for a digital repository and archival system. The OAIS reference model provides a common language and conceptual framework that digital library and preservation specialists now share. The framework has been adopted as an International Standard, ISO 14721:2003. Though some critics identify shortcomings in the detail of the OAIS, the concept of constructing repository architectures in a form that corresponds with the OAIS functional categories is critical to the development of modular storage systems with interoperable exchange of content. The following sections of the Guidelines adopt the major functional components of the OAIS reference model to assist in the analysis of the available software and to develop recommendations for necessary development.

6.1.8.2 There are a finite number of functions an archival digital repository must be able to perform in order for it to reliably and sustainably perform the purpose for which it is designed. These are defined in the Reference Model for an Open Archival Information System (OAIS) as Ingest, Access, Administration, Data Management, Preservation Planning and Archival Storage.



6.1.8.3 The OAIS also defines the structure of the various information packages that are necessary for the management of the data, according to the place in the digital life cycle. These are the Submission Information Package (SIP), Dissemination Information Package (DIP) and Archival Information Package (AIP). A package is the conceptual parcel of the data and relevant metadata and descriptive information necessary to the particular object. This object is conceptual only in the sense that the package contents may be dispersed in the system or collapsed into a single digital object. OAIS defines an information package as the Content Information and associated Preservation Description Information which is needed to aid in the preservation of the Content Information.

6.1.8.4 The SIP is an Information Package that is delivered to the system for ingest. It contains the data to be stored and all the necessary related metadata about object. The SIP is accepted into the system and used to create an AIP.

6.1.8.5 The AIP is an Information Package which is stored and preserved within the system. It is the information package the system stores, preserves and sustains.

6.1.8.6 The DIP is the information package created to distribute the digital content. There are three roles in this system. First is access, and this DIP would be in a form that the users can use and understand. Second is exchange for the purpose of distributing risk. An archival repository may choose to share parts of its content with other similar institutions, or with an organisation whose role is archival storage. In this case the DIP would contain all the relevant metadata necessary to undertake this role. The third is for distributing content to archives as a last resort. The scenario where a particular archive or institution no longer has the resources to maintain its collection is not difficult to imagine. A standard DIP for this purpose allows other similarly architected systems to undertake the role with the minimum of manual intervention.

6.1.9 Trusted Digital Repositories (TDR) and Institutional Responsibility

6.1.9.1 The technical specification of the digital storage environment is an important part of ensuring that the digital content that is managed is still accessible to researchers in the future. It is not of its own, however, enough to ensure that this will be achieved. The institution within which the digital archive resides has to be able to ensure that the content it manages is curated and maintained responsibly. In 2002 the Research Libraries Group (RLG) and the Online Computer Library Center (OCLC) jointly published "Trusted Digital Repositories: Attributes and Responsibilities" (TDR), which articulated a framework of attributes and responsibilities for trusted, reliable, sustainable digital repositories which were "required for an archive to provide permanent or indefinite long-term preservation of digital information".

6.1.9.2 These attributes include compliance with the OAIS reference model, organisational viability, financial sustainability, technological and procedural suitability, the security of the system and the existence of appropriate policies to ensure that the steps are taken to manage and preserve the data.

6.1.9.3 The practical instantiation of this is a document known as the "Trustworthy Repositories Audit and Certification (TRAC): Criteria and Checklist" (2007). Using this document an archival institution can establish whether the practices, approaches and technologies they have or are planning to implement are appropriate to the permanent preservation of the digital information for which they have responsibility.

6.1.9.4 The concern which the checklist addresses incorporates three main areas: organisational infrastructure; digital object management and technologies; and technical infrastructure and security.

6.1.9.5 Organisational infrastructure provides a series of checks against appropriate governance and organisational viability, organisational structure and staffing, procedural accountability and policy framework, financial sustainability and a consideration of the licenses, and liabilities. Digital object management section considers the acquisition of content, the creation of an archivable package, planning for preservation, archival storage and planning, information management and access control. The third part of this checklist audits the system infrastructure, the use of technologies appropriate to the tasks and system and institution security.

6.1.9.6 The terminology used in the "Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist" is chosen to represent digital archives in the broadest sense of the word, and so the document's meaning may occasionally appear opaque to an audio archivist. Nonetheless, the issues examined and tested by it are critical to the planning and management of a digital audio archive. It is strongly recommended that the digital sound archivist uses the checklist to examine the suitability of an institution to manage a digital collection, or to identify weaknesses within an existing digital preservation strategy.

6.1.10 Audio Archives and Technical Responsibility

6.1.10.1 Though a particular institution may be responsible for the management of a collection or set of audio items, it does not necessarily follow that institution will undertake the responsibility for maintaining the digital storage system. An institution may instead become a part of a distributed storage system, or may identify a third party provider to archive their content in a more standard approach.

6.1.10.2 A distributed data storage approach such as that being promoted and developed for web based material by Stanford University under the name of LOCKSS (Lots of Copies Keep Stuff Safe) replicates data in a number of places on the web. The system manages the data on the grid and risk of loss of data is reduced because the information can be found in many different places. Such a system is not appropriate for material which has access restrictions or copyright which prohibits dissemination. Such a system also requires that a development and management responsibility to be shouldered by an institution.

6.1.10.3 An institution may decide that they do not have the technical capability to undertake the development and management of a digital storage system. In this case they may establish a relationship with a third party provider. That provider may be another archive which will take the collection and store its content, or may be a commercial provider who will provide and manage the storage and content for a fee.

6.1.10.4 The information provided here is provided as though the institution is intending to take on its own preservation. However, if any of the above alternatives are considered, then this information is useful for determining if those approaches are reliable and valid.

6.1.11 Digital Repository Software, Data Management, and Preservation Systems:

6.1.11.1 Digital repository software is generally that software which supports storage and access to the digital content. It should incorporate indexing and metadata systems that manage information about the content, and a variety of tools to find and report on the content.

6.1.11.2 Data management is the management of the byte stream, or data, that the system is responsible for. This may include back up procedures, multiple copies and changes.

6.1.11.3 Preservation processes are those that ensure the content remain accessible in the long term, that the content is still meaningful and that the data management system's tasks are documented and maintained. All three of these steps are necessary to achieve long term preservation to content.

6.2 Ingest

6.2.1 Submission Information Package (SIP)

6.2.1.1 The SIP is an Information Package that is delivered to the repository and digital storage system for ingest. The SIP includes the audio data to be stored and all the necessary related metadata about the object and its content. Ingest, in the OAIS model, is the process that accepts the content and all its related metadata (SIP), verifies the file, extracts the relevant data and prepares the AIP for storage, and ensures that AIPs and their supporting Descriptive Information become established within the OAIS.

6.2.1.2 A digital repository and preservation system should be able to accept and validate an audio file. Validation is a process that ensures that the files which are being accepted into the digital storage system comply with the standards. Non standard files may become difficult to use in the future when current replay systems no longer exist. Tools exist for automated validation of file formats, and some open source solutions, like JHOVE (JSTOR/Harvard Object Validation Environment), are available and being further developed.

6.2.2 Format

6.2.2.1 IASA recommends the use of .wav or preferably BWF .wav files [EBU tech 3285]. The difference between the two is that the BWF contains a set of headers which can be used to organise and manage metadata. Though BWF metadata is adequate for many purposes, in some sophisticated systems and exchange situations a more comprehensive package is required, and in these circumstances Metadata Encoding and Transmission Standard (METS) is often used. The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using XML (eXtensible Markup Language). A METS package, which consists of metadata and content, is often used as an exchange standard between digital archives.

6.2.2.2 Material eXchange Format (MXF) is a container format for professional digital video and audio media defined by a set of SMPTE standards. MXF has been mostly taken up by the video archiving community, though it is capable of managing audio. Like METS, it is primarily a set of metadata which "wraps" the content, in this case, audio. Both these are very useful formats in the exchange and management of content and information between archives and repositories.

6.2.2.3 The format of the SIP will depend on the system and the size and sophistication of the enterprise. It is quite possible to establish a viable archive using .wav files and manually entering most of the necessary metadata into the system by hand, and acquiring the necessary technical metadata at the ingest stage. This however, would only be appropriate for the smallest of collections. Large collections with remote and separate digitisation processes and large quantities of material must build sophisticated ingest and data exchange systems to ensure the content is adequately ingested into the data storage systems. Production and verification software generates much of this data as standardised XML-files that may be used for preservation purposes. The National Library of New Zealand Metadata Extractor tool, for example, is a Java-based tool that extracts preservation metadata from digital objects and outputs that metadata in a standard format (XML).

6.2.3 Preservation Metadata

6.2.3.1 The metadata needed to manage preservation processes at the ingest stage is all the information regarding the creation of the digital audio object and the changes to format that have occurred prior to ingest. In this way the technical provenance of the object is preserved, which allows a pathway between the present form of the item and original from which it was created to be traced.

6.2.3.2 BWF has a non-compulsory recommendation for BWF entitled "Format for CodingHistory field in Broadcast Wave Format" http://www.ebu.ch/CMSimages/en/tec_text_r98–1999_tcm6-4709.pdf which describes how changes to the file may be described. Local usage of the ASCII free text field allows the description of the technical equipment or software that was used in the creation of the digital audio object.

6.3 Archival Storage

6.3.1 Archival Information Package (AIP)

6.3.1.1 The definition of the term Archival Storage in OAIS includes the services and functions necessary for the storage of the Archival Information Package (AIP). Archival storage encompasses data management and includes processes such as storage media selection, transfer of AIP to storage system, data security and validity, backup and data restoration, and reproduction of AIP to new media.

6.3.1.2 AIP, as defined in OAIS reference model (CCSDS 650.0-B-1 Reference Model for an Open Archival Information System (OAIS)), is an information package that is used to transmit archival objects into a digital archival system, store the objects within the system, and transmit objects from the system. An AIP contains both metadata that describes the structure and content of an archived essence and the actual essence itself. It consists of multiple data files that hold either a logically or physically packaged entity. The implementation of AIP can vary from archive to another archive; it specifies, however, a container that contains all the necessary information to allow long term preservation and access to archival holdings. The metadata model of OAIS is based on METS specifications.

6.3.1.3 From physical point of view the AIP contains three parts; metadata, essence and packaging information, which all consists of one or more files (see 6.1.3 Defining the Digital Object). Packaging information can be thought as wrapper information and it encapsulates metadata and essence components.

6.3.2 Archival Storage basics

6.3.2.1 Archival Storage provides the means to store, preserve and provide access to archived content. In small systems the storage can stand alone and may be manually operated, but in larger systems storage is usually implemented in conjunction with cataloguing applications, asset management systems, information retrieval systems and access control systems in order to control and manage archived content and provide a controlled way to access them.

6.3.2.2 Archival Storage must be connected to equipment that ingests and creates the digital asset to be archived, and it must provide a secure and reliable interface that can be used to import assets to the storage system.

6.3.2.3 A system that is used to store archival content must be reliable in several ways: It must be available for use without any significant interruptions, and it must be able to report to the system or user who imports content whether the import was successful or not, thus enabling the importing party to delete the ingest copy of the of the archival file if appropriate. Archival Storage must also be able to preserve the content it manages for a long period of time and be able to protect the content from all kinds of failures and disasters.

6.3.2.4 An Archival Storage system should be built according to the needs of its functional owner: it must be correctly-sized to carry out the tasks that are needed, and manage the capacities that are required in every day operations. In addition, Archival Storage must provide controlled access to the content it manages for the users who have permissions or rights to access the content.

6.3.3 Digital Mass Storage Systems (DMSS)

6.3.3.1 A Digital Mass Storage System refers to an IT based system that has been planned and built to be able to store and maintain large amounts of data for a given or extended period of time. These systems come in many forms; a basic DMSS could be a personal computer which has large enough hard disk drive and some kind of catalogue that can be used to keep track of the assets the system possesses.

A more complex DMSS may consist of hard disk drive and/or tape storage and group of computers that control the storage entity. A DMSS can also contain many tiers of storage with different characteristics; a fast Fibre Channel based hard disk drive tier can be used to cache assets whose access time is critical while a tier built of cheaper hard disk drives could be used to hold material whose access time is not so critical, and finally tape based storage can be used as the most cost-effective tier of storage.

6.3.3.2 When a number of different storage technologies are used in a large system to build the functional entity, a HSM (Hierarchical Storage Management) system is usually deployed in such a way that it supports the different technologies working together. Larger scale systems may also be distributed geographically in order to achieve better performance and make the system more fault tolerant.

6.3.4 Data Tape Types and Formats Introduction

6.3.4.1 The following is an outline of some of the main data tape formats and tape automation systems that may be used for storing AV content in data form. Data tapes are only used in conjunction with other components of a DMSS. It is prudent to commence a section on comparison of the various data tape formats with a reminder that no carrier is permanent and that, all things being equal, they will only be viable as long as the data systems in which they are incorporated continue to support them.

6.3.5 Data Tape Performance

6.3.5.1 Format geometry and dimensions govern performance. Data transfer speed, one aspect of performance, is a direct product of the number of tracks written and read simultaneously, as well as the tape-head speed, linear density and the channel-code. Similarly, physically smaller, lighter tape housings are faster to move in a robotic library. Data density is a product of:

6.3.5.1.1 tape length and thickness trade-offs

6.3.5.1.2 track width and pitch

6.3.5.1.3 linear density of data payload within each track

6.3.6 Tape Coatings

6.3.6.1 There are two main types of tape coatings: particulate and evaporated. The earliest coated data tapes used metal oxides similar to video tape, whereas more recent data tapes use metal particles (MP). Pure iron with inert ceramic and oxide passivation layers is dispersed in polymer binders which are applied evenly to a PET or PEN base-film or substrate which in turn provides dimensional stability and strength under tension. Some of the highest density data tapes currently on the market now use evaporated metal foil coatings of cobalt alloys and similar material to those used on hard disks. This achieves a much higher purity of magnetic material and allows thinner coatings. Most metal-evaporated (ME) tapes have a protective polymer coating similar to the binder material on MP tapes. The more recent formulations include a ceramic protective layer as well. Several of the early ME tapes failed during heavy usage due to de-lamination (Osaki 1993:11).

6.3.7 Tape Housing Design

6.3.7.1 Two basic styles of housings are used, dual-hub cassettes, which may enable faster access times and single-hub cartridges, which offer greater capacity within a given external volume.

6.3.7.2 Dual hub cassettes include:

3.81mm, principally DDS [derived from DAT]

QIC [quarter-inch cartridge] and Travan
8mm formats, including Exabyte and AIT
DTF
Storagetek 9840

6.3.7.3 Single-hub cartridges include:
IBM MTC and Magstar formats such as 3590, 3592 and TS1120
Quantum S-DLT and DLT-S4
LTO Ultrium [100, 200, 400 & 800 GB]
Storagetek 9940 and T10000
Sony S-AIT

6.3.7.4 Neither design is necessarily superior for long-term archiving, since the life is governed by a range of details specific to each format. For instance, some models of the single-ended ½-inch cartridges have large-diameter guides within the housing, which ensure minimum friction and accurate tape guidance. Problems have been experienced with the leader latching mechanism on older single-ended cartridges, although more recent designs have improved reliability in this area. Some dual-hub cassettes can be positioned to park halfway along the tape to minimise the amount of spooling time to any particular file. This contradicts the traditional practice in AV archives of spooling tapes carefully to one end before storage so that only leader tape is exposed to the threading mechanism. Tapes generally don't incorporate a hermetically sealed enclosure in the way that hard disks are protected.

6.3.8 Linearly and Helically Scanned Tapes

6.3.8.1 Data tapes may be written or read with a fixed head, generally described as linear, or with a rotating or helical head. Linear tapes typically follow a serpentine track layout, and it has been argued that this shuttling can lead to wear or a so-called shoe-shine effect. In practice, modern tapes are designed to last for large numbers of passes, however, it is still prudent to access frequently used content from hard disc. Tapes, which experience chemical decomposition from hydrolysis and other causes, will usually run better over fixed guides and components in the tape path at speeds of around1-2 m/s or greater, which are typical of fixed-head or linear formats. Rotary-head or helical formats typically have higher tape-head speeds which create a greater air-bearing effect between the tape surface and the read-write heads, but the linear tape speed over the fixed guides and heads is much slower, so this is where fouling often occurs.

6.3.9 Ancillary Storage and Access Devices

6.3.9.1 Formats such as AIT include solid-state 'Memory in Cassette' or MIC which stores file positional information similar to a Table of Contents (TOC) on Compact Disks for rapid location of data. DTF uses rf memory.

6.3.10 Format Obsolescence and Technology Cycles

6.3.10.1 The inherent nature of data storage is of constant progress and development, which means inevitable change, and ongoing obsolescence. Realistic long-term management of content must accept and build upon the continuing evolution and upgrading of hardware and media. Although central infrastructure such as data cabling or storage libraries may remain in operation for ten or twenty years, individual tape drives and media have a finite life much shorter than this. All of the main data tape formats have development roadmaps projecting upgrades every 18 months to 2 years. Backward compatibility for read-only access is sometimes assured over one or two generations of media within any common family. As a result, each generation of tape drives and media may be viable for 4 to 6 years, after which time it is essential to migrate the data and move on.[1] Also the hardware maintenance cost of mass storage systems tend to rise notably when the system gets older than its projected life or the guarantee period ends. After this it may be difficult to obtain new spare parts for the tape libraries or tape drives, for example. A summary of projected roadmaps is presented below. Many formats have read-only compatibility with at least one previous generation.

Family	1st Generation	2nd Generation	3rd Generation	4th Generation	5th Generation	6th Generation
Quantum SDLT	SDLT220 110GBytes	SDLT320 160GBytes	SDLT600 300GBytes	DLT-S4 800GBytes		
IBM			3592 2004 300GB 40MB/s	TS1120 2006 700GB 104MB/s		
Sun - Storagetek		9940B 2002 200GB 30MB/s	T10000 2006 500GB 120MB/s	T10000B-2008 1TB 120MB/s		
LTO	LTO-1 2001 100GB 20MB/s	LTO-2 2003 200GB 40MB/s	LTO-3 2004 400GB 80MB/s	LTO-4 2007 800GB 120MB/s	LTO-5 no date (2009+) 1.6TB 180MB/s (estimated)	LTO-6 no date (2011+) 3.2TB 270MB/s (estimated)
Sony S-AIT	S-AIT 2003 500 GB 30MB/s	S-AIT2 2006 800 GB 45MB/s				
Sony AIT			AIT-3 2003 100 GB 12MB/s	AIT-4 2005 200 GB 24MB/s		

Table 1 Section 6.3: Projected Development Roadmap for Data Tapes

6.3.11 Automated Robotics or Manual Retrieval

6.3.11.1 For small-scale operations it is possible to back up data from a single workstation onto a single data tape drive and manually load tapes for storage on traditional shelving, and even small scale networked systems will undertake manual backup of their storage (see also Chapter 7 Small Scale Approaches to Digital Storage Systems). The same guidelines for storage environments apply as for other magnetic tapes, though increased attention to minimising the presence of dust and other particulates and pollutants would be beneficial. For larger-scale operations, particularly in countries where labour costs are high, and capital equipment budgets are favourable, a degree of automation is normally desirable and more economical than purely manual systems. The degree of automation depends upon the scale and consistency of the task, type of access to the content, and the relative costs of the main resources.

6.3.11.2 **Autoloaders and Robotic Tape Libraries:** The next step from single drives is the small-scale autoloader, which usually has one drive (occasionally two), and a single row or carousel of data tapes which are fed in sequentially to support backup operations. One of the key differences between autoloaders, and large-scale robotic libraries is that the recorded tapes are not logged by the backup software in a central database which can then enable automated retrieval. The task of searching,

[1] This implies a degree of waste and environmental pressure beyond the scope of our purely technological discussion, but in reality, a large-scale library of older data tapes will consume more polymers and require more petrochemicals for manufacture than a newer, high-density system with more energy-efficient drives and robotics, occupying less real-estate at the same time.

retrieving and reloading individual files still falls to a human operator. All that autoloaders do, as the name implies, is to allow a series of tapes to be written or read sequentially to overcome the size limitations of individual data media, and to negate the requirement for a human operator's presence to load the next tape in a long backup sequence.

6.3.11.3 By way of contrast, even the smallest robotic tape libraries are programmed to behave as a single, self-contained storage system. The location of individual files on different tapes is transparent to the user, and the library controller keeps track of addresses of files on each tape, and of the physical location of tapes within the library. If tapes are removed or reloaded, the robotic sub-system re-scans the tape slots as it initialises, to update its inventory with metadata from barcodes, rf tags, or memory chips in the tape housings.

6.3.11.4 Large tape libraries have some benefits when compared to the smaller tape libraries. They can be built to be redundant and distributed, i.e. downtime can be minimised and the read/write load can be balanced between several similar systems. Large tape library can also be used as a multi-purpose system; they can, for example, maintain a company's normal IT backups as well as manage all archived video and audio.

6.3.11.5 Data tapes or cartridges used in a robotic system will have some system of barcoding, rf tags or other ID. These optical or electromagnetic recognition systems sometimes operate in conjunction with MIC for supplementing information about tape ID and content. Some formats have a global ID system for barcoding tapes so that a tape used in one robotic library can be recognised in another library system.

6.3.11.6 **Backup and Migration Software and Schedules**: Some confusion and misunderstanding exists both in IT circles, and in the wider community as to the purpose and operation of long-term data archives. There are two popular misconceptions regarding long term data archives. The first; that archiving is the process of moving infrequently used material from expensive, on-line networked disc storage, to less expensive, inaccessible offline shelving from whence it may never be retrieved and the other; that backup is a regular daily and weekly routine of making a copy of everything stored in the system.

6.3.11.7 With regard to the first misconception, the reality is that some of the most important and valuable material may not be used for months or years, but its survival must be guaranteed unequivocally. Likewise with the second, if suitable rules are established, vast amounts of material may not need to be replicated daily or weekly when only small percentages are updated. In practice, while a stringent regime of replicating data on different media in different locations is essential to minimise risks from technology failures and to ensure recovery from disasters, the particular characteristics of digital heritage material requires some procedures that differ from routine IT data management.

6.3.11.8 Conventional HSM (Hierarchical Storage Management) systems may be optimised for backing up everything on a regular basis, and moving out infrequently-used content to inaccessible locations, but the better systems can be configured to suit the business rules and practices in archives of different sizes with different levels of access. A medium-sized organisation may ingest 100 GB of audio data every week or 1TB of video. It is fairly straightforward to ensure that copies are made as soon as valuable material is ingested, and that frequently used material remains accessible.

6.3.11.9 Some of the primary tasks of storage management software are to optimise the use of resources and to manage devices in the hardware layer, while regulating traffic with minimal delays to users. HSM software offers a choice of conditions for migrating files from on-line disk to tape, such as older than a certain date, larger than a nominated size, located in particular sub-folders or when available disk space falls outside certain limits (high and low watermark).

6.3.11.10 Typically, where both high resolution files, as well as low resolution access copies are produced, the larger, high resolution files used for preservation and broadcast will be migrated to tape to free up space on the more expensive hard disk array. A balance is needed to maintain availability of material, and to optimise use of tape drives and media. If tapes are being accessed very frequently, a large number of mounts and unmounts, spooling and restore operations will degrade system performance. More sophisticated content management systems sometimes incorporate lower levels of storage management so that users are less aware of individual files and components that support the system.

6.3.12 Selection and Monitoring of Data Tape Media

6.3.12.1 As with any conventional preservation system, it is important not only to have backups and redundancy in case of failures in media or components, but it is vital to establish and to measure performance standards for key parts of the system. Software such as SCSI-Tools will allow a lower level of interrogation of individual drives and devices on a network to determine if media and hardware are performing at their optimum level. LTO tape has an interface for data monitoring, however this functionality is rarely utilised though it would be advantageous for archival systems. Some HSM systems are capable of monitoring the quality of stored assets on a regular basis. These systems monitor the error rates of tapes while users access the assets or read the assets without user intervention if a tape has not been used during a certain period of time.

6.3.13 Costs

6.3.13.1 Typically, the cost of data tape storage is spread in four areas: Tape media: procurement and replacement of primary and backup tape media every 3–5 years. Tape drives: procurement and replacement every 1–5 years, with support. Robotic Library purchase and maintenance within a 10 year life-cycle, and software purchase, integration/development and maintenance.

6.3.13.2 In a manual system, the costs for shelving will be lower, although the space requirement for staff is greater, and the labour cost for manual retrieval and checking is higher. In an automated robotic system, much of the human cost is offset by up-front expense for hardware and software. Large scale robotic tape libraries can be purchased in a modular fashion to spread the cost over several years as demand for storage grows. Within the life of a robotic tape library, individual components such as tape drives will be replaced by newer technology every three to five years. If content from an archive is accessed continuously the life time of drives can be considerably short, even only one year or less. Older tape media and drives may be kept on hand for redundancy if required. If an archive does not grow rapidly, the present and next generation of tapes and drives can co-exist in a tape library while the archive content is migrated to the next generation of media or technology. If an archive grows continuously it may be cost-effective to create a tape library of a specific size to only store the amount of content that shall be archived during the life time of the then current technology, and to then acquire a new larger tape library to store the content that shall be stored using the next generation of technology including the old content that will be migrated. The later approach is also necessary if old and new technology cannot co-exist in the same unit.

6.3.13.3 It is good business practice to keep at least one redundant copy of data off-site or geographically separate. Typically a radius of 20 to 50 km is common for natural and man-made disasters, and still allows manual retrieval within a few hours. To reduce risks further, redundant copies should be on different batches or sources of media, or even on different technologies. Some data tapes are only manufactured at a single supplier, and chances of a single point of failure are increased. Three copies of data are safer than two, and although costs for media increase, the hardware and software costs are only slightly higher than for the first copy.

6.3.14 Hard Disk Drives (HDD) Introduction

6.3.14.1 Hard Disk Drives (HDDs) have served as the primary memory and data storage in computers since IBM introduced the model 3340 disk drive in 1973. Nicknamed "the Winchester", because it had 30MB of fixed memory and 30MB of removable and the working designation of 30/30 resembled, in name at least, the famous rifle, it pioneered head designs that made operation of the hard disk viable. Subsequent reduction in size and more recent developments in head and disk design have greatly increased the reliability of disk drives, leading to the robust designs in common use today.

6.3.14.2 Data managers whose responsibility it is to maintain data have considered the hard disk too unreliable to use as the sole copy of an item, and too expensive to use in multiple, and consequently more reliable, disk arrays. The data on HDDs has consequently been duplicated on multiple tape copies to ensure its survival. As stated above (6.1.4 Practical Aspects of Data Protection Strategies and 7.6 Archival Storage) all data systems must have multiple and separate copies of all data. While experts tend to agree that the most reliable data system consists of a HDD array supported by multiple duplicates on tape, the continued reduction in costs and improvement in reliability make the concept of identical duplicates of data on separate hard disks a possibility. The principle of multiple media remains, however, and disk only storage constitutes a risk.

6.3.15 Reliability

6.3.15.1 Loss of data as a consequence of disk failure and head crashes has made most data professionals suspicious of HDDs, however manufacturers now claim annualised failure rates of less than one percent and an operational life of 40,000 hours (Plend 2003). High reliability drives may have an even longer operational life, termed by manufacturers as "mean time between failure". Though HDDs are self-contained and sealed and so protected from damage, most failures in disk drives occur in two opposing ways: as a result of wear through extended use, or as power to the drive is turned on or off. The dilemma is whether to leave the disk on, and increase wear, or turn it on and off and increase risk of failure.

6.3.16 System Description, Complexity and Cost

6.3.16.1 As noted in Section 2, Key Digital Principles, the more recent generations of computers have sufficient power to manipulate large audio files. All recent generation computers incorporate hard disks of adequate speed and size, and an external HDD adapter can be plugged into a USB, Firewire or SCSI port. The system complexity and the degree of expertise required to run such systems is not much greater than is necessary for desktop computer operation.

6.3.16.2 When large quantities of audio and audiovisual material required for access are stored on HDDs, the disks are usually incorporated into a RAID (Redundant Array of Inexpensive (or Independent) Disks). RAID increases the reliability of the hard disk system, and the overall access speed by treating the array of disks as one large hard disk. If a disk fails, it can be replaced and all the data on that disk can be reconstructed with data from the rest of the disks in the array. The level of failure the system will tolerate, and the speed of recovery from such failures is a product of the RAID levels. RAID is not designed as a data preservation tool, but as a means of maintaining access through inevitable disk failures. The appropriate RAID level for any particular installation, and the requirement for duplication of controllers, is dependant on the particular circumstance and the frequency of data duplication. A RAID requires that all disks in the array be turned on when any part of the disk is in use. All RAIDs containing archival material, as with all digital data, must be duplicated more than once on other carriers.

Capacity	Native tape capacity (GB)	# of tapes	Recommended # of tape drives	Maximum # of drives	System price (€)	Tape price (€)	Drive price (€)	Cost per GB (€)
10 TB	800	13	2	4	20.480	97	7.625	2,05
50 TB	800	63	4	16	56.800	97	10.175	1,14
100 TB	800	125	8	16	134.050	97	12.725	1,34
200 TB	800	250	12	16	205.350	97	12.725	1,03
500 TB	800	625	18	56	446.938	97	15.975	0,89
1000 TB	800	1250	36	88	864.517	97	15.975	0,86
2000 TB	800	2500	72	176	1.687.690	97	15.975	0,84

Table 2 Section 6.3: Investment Costs of LTO-4 technology based Storage Systems

Capacity	HW maintenance, year 1 (€)	SW maintenance, year 1 (€)	HW maintenance, year 2 (€)	SW maintenance, year 2 (€)	HW maintenance, year 3 (€)	SW maintenance, year 3 (€)	HW maintenance, year 4 (€)	SW maintenance, year 4 (€)	HW maintenance, year 5 (€)	SW maintenance, year 5 (€)
10 TB	2.420	n/a	2.420	n/a	2.420	n/a	2.514	n/a	2.514	n/a
50 TB	3.454	n/a	4.958	n/a	4.958	n/a	4.958	n/a	4.958	n/a
100 TB	11.808	490	13.817	490	13.817	490	13.817	490	13.817	490
200 TB	15.787	582	19.323	582	19.323	582	19.323	582	19.323	582
500 TB	27.380	1.068	34.111	1.068	34.111	1.068	34.111	1.068	34.111	1.068
1000 TB	47.542	2.115	66.734	2.115	66.734	2.115	66.734	2.115	66.734	2.115
2000 TB	99.272	4.221	99.272	4.221	99.272	4.221	99.272	4.221	99.272	4.221

Table 3 Section 6.3: Yearly Maintenance Costs of LTO-4 technology based Storage Systems

Notes to the tables:

- Prices are averages of list prices from multiple vendors. A price that a customer has to pay is usually somewhat lower.
- Prices indicate price of raw capacity. At least double amount of tape media will be needed for backup purposes.
- Price in the system price column includes cost of tapes and drives for the capacity in question, but does not include any HSM software or hardware
- The tables indicate only investment costs and maintenance fees that have to be paid to a vendor. In addition to this, also costs from electricity, cooling, machine room, management, etc. must be included in individual calculations. Electricity and cooling of tape library system might cost 10% of purchase price over five year period.

Capacity	Drive technology	Size of drive (GB)	# of drives	System price (€)	Drive price (€)	Cost per GB (€)
5 TB	SATA	500–1000	5–10	11.884	1.000	2,38
10 TB	SATA	750–1000	10–14	19.997	1.000	2,00
50 TB	SATA/FATA	1000	50	124.334	1.800	2,49
100 TB	SATA/FATA	1000	100	230.914	1.800	2,31
200 TB	SATA/FATA	1000	200	456.942	1.800	2,28
500 TB	SATA/FATA	1000	500	1.202.726	1.900	2,41
1000 TB	SATA/FATA	1000	1000	2.566.513	1.900	2,57
2000 TB	SATA/FATA	1000	2000	4.782.584	1.900	2,39

Table 4 Section 6.3: Investment Costs of HDD Based Storage Systems

Capacity	HW maintenance, year 1 (€)	SW maintenance, year 1 (€)	HW maintenance, year 2 (€)	SW maintenance, year 2 (€)	HW maintenance, year 3 (€)	SW maintenance, year 3 (€)	HW maintenance, year 4 (€)	SW maintenance, year 4 (€)	HW maintenance, year 5 (€)	SW maintenance, year 5 (€)
5 TB	826	750	826	750	826	750	1.845	750	1.845	750
10 TB	1.206	1.125	1.206	1.125	1.206	1.125	2.600	1.125	2.600	1.125
50 TB	5.822	6.125	5.822	6.125	5.822	6.125	12.365	6.125	12.365	6.125
100 TB	10.514	8.500	10.514	8.500	10.514	8.500	22.391	8.500	22.391	8.500
200 TB	21.724	12.750	21.724	12.750	21.724	12.750	44.956	12.750	44.956	12.750
500 TB	57.061	37.250	57.061	37.250	130.394	37.250	130.394	37.250	130.394	37.250
1000 TB	130.203	66.250	130.203	66.250	263.537	66.250	263.537	66.250	263.537	66.250
2000 TB	223.778	124.250	223.778	124.250	477.121	124.250	477.121	124.250	477.121	124.250

Table 5 Section 6.3: Yearly Maintenance Costs of HDD Based Storage Systems

Notes to the tables:

- Prices are averages of list prices from multiple vendors. A price that a customer has to pay is usually somewhat lower.
- Price in the system price column includes cost of hard disk drives for the capacity in question.
- The tables indicate only investment costs and maintenance fees that have to be paid to a vendor. In addition to this also costs from electricity, cooling, machine room, management, etc. must be included in individual calculations. Electricity and cooling of hard disk drive system might cost 30% to 40% of purchase price over five years period.

6.3.17 Disk Only Storage

6.3.17.1 RAID arrays are scalable within the limits of the system, however individual HDDs are infinitely scalable by simply adding more drives. Since the introduction of the IBM 3340 HDD, storage capacity has increased rapidly, almost exponentially, while costs have fallen. These changes, linked with an improvement in reliability, have led some to suggest that HDDs could be used as both the primary storage system and the back up copy. There are three difficulties associated with this approach: Firstly, hard disk life is estimated in terms of usage-time, that is the number of hours of operation. There has been no testing of the life of an infrequently used HDD. Secondly, having data on different types of media is advantageous as it spreads the risk of failure. Therefore the approach should be considered very cautiously. Finally, there is no way of monitoring the condition of the hard disk on the shelf without turning it on at regular intervals and thereby compromising the advantage gained by having the disk turned off (see section 6.3.18 below, Monitoring of Hard Disk Media). Multiple carriers (eg Tape and Hard disk) remain the preferred option. Hard disks should be implemented within an integrated system.

6.3.18 Hard Disk Storage Systems

6.3.18.1 Hard Disk Storage Systems are centralised systems that are used to maximise disk storage utilisation and to provide large capacities and/or high performance. These systems are used in conjunction with server computers so that server have only small amount of internal hard disk storage or do not have it at all. These kind of systems are often used in mid and large size environments as storage for an archiving system. Alternatively an archiving system can share a centralised storage system with a number of other computer systems. The size of a system can vary from 1 terabyte to several petabytes. It should be taken into consideration that performance characteristics of a storage system can vary notably according to its chosen configuration and it is essential that the actual needs for a system are carefully planned beforehand and a qualified professional is used to configure the storage structure and interfaces of a system to produce the best value for ones investment.

6.3.18.2 Centralised disk storage systems are designed to provide better error resilience than independent hard disk drives. These systems provide several alternative levels of RAID protection, their components can be redundant in order to avoid single point of failures, and systems can be locally or geographically distributed to protect valuable assets from different kind of failures and disasters.

6.3.18.3 The connection between the storage system and the computers it serves play important role regarding performance of a system. Generally speaking, two methods used are NAS (Network Attached Storage) and SAN (Storage Area Network). While NAS utilises regular IT network like Ethernet to move data between computer and storage system SAN uses switched Fibre Channel connections. NAS systems can operate at 100 Mbit/s, 1 Gbit/s and 10 Gbit/s speeds while SANs operate at 2 Gbit/s or 4 Gbit/s. Both technologies have clear road map to the future and their performance can be expected to grow in the future. SAN technology is usually chosen for more demanding environments since it gives better performance due to specific design. For example, the in/out (I/O) block size can be controlled more effectively in SAN environments while networking protocols tend to force NAS systems to use quite small I/O blocks. From economical point of view NAS technology is cheaper than SAN technology.

6.3.19 HDD Life

6.3.19.1 As stated above, a life of 40,000 hours is estimated for many commercially available HDDs. Typical commercial use of HDDs would give these disks a replacement life of five years. With improvements such as fluid/ceramic spindle bearings, surface lubrication of disks, and special head parking techniques made on the most recent desktop HDDs, the life of HDDs may be somewhat longer. However, there is no reliable testing of the life span of unused HDD and it would be astute to plan to replace the disks in such a working system within 5 years.

6.3.20 Monitoring of Hard Disk Media

6.3.20.1 An indication of imminent disk failure may be an increase in bad data blocks. It is typical for the latest disks to show bad block errors even from new and most data systems manage the bad blocks by reassigning the address of that block. However, if the quantity of bad blocks increases it may indicate that the disk is beginning to fail. Software exists which will provide a warning of increased bad data blocks, as well as measuring other physical characteristics that may indicate disk problems.

6.3.21 HDD technologies

6.3.21.1 There are four main methods of connecting HDDs and other peripheral devices to computers, USB (Universal Serial Bus), IEEE 1394 (Firewire), SCSI (Small Computer System Interface) and SATA/ATA (Serial Advanced Technology Attachment/AT Attachment). They each have particular advantages in certain situations. USB and Firewire are planned to be all-purpose buses that can be used to connect to personal computer a HDD as well as digital video camera or MP3 player. SCSI and SATA/ATA are mainly used to connect hard disk drives to a computer or disk storage system.

6.3.21.2 SCSI and its successor SAS (Serial Attached SCSI) interface allows faster writing and reading speeds, and facilitates access to larger numbers of drives than the SATA/ATA drives. SCSI disks can accept multiple commands at once on a SCSI bus and does not suffer from request queues like SATA/ATA. The SATA/ATA drives are comparatively cheaper. The read access speed is largely the same and in an audio context neither interface will limit the operation of the digital audio workstation (DAW) more than the other. The performance difference of SCSI/SAS and SATA drives can have meaning in heavily utilised centralised hard disk storage system.

6.3.21.3 Fibre Channel (FC) SCSI/SAS drives are mainly used in demanding use in enterprise or business systems while the cheaper SATA drives are more used in the personal market, but they are also increasingly used in enterprise and business systems to offer more cost-effective storage capacity e.g. in archival storage. In archival storage, the actual decision between (FC) SCSI/SAS and SATA technology is dependent on the actual load of the system. If a system is used to archive small or medium amounts of content that is not accessed intensively a SATA based solution might well be enough. The actual decision must be based on clearly identified demands and negotiations with one's storage provider.

6.3.21.4 USB and Firewire connected disk can be used to transfer content from one environment to another, but since they are rather unreliable, difficult to monitor and easy to loose they should not be used for archiving even though their pricing may seem very attractive.

6.3.21.5 The interface is not a completely consistent indication of the reliability and performance of a given drive or storage system and the purchaser should be more aware of other operating and configuration parameters of a storage system. It seems to be the case that more reliable drives are associated with the FC SCSI/SAS interface. Nonetheless, HDDs are not in themselves permanently reliable, and all audio data should be backed up on suitable tape (see 6.3.5 Data Tape Performance). (For further discussion see Anderson, Dykes and Riedel 2003).

6.3.21.6 There is one emerging storage technology which may have a prominent position in the near future. Solid-state storage in form of flash memory is developing as a alternative to moving disks and has already become an alternative to a HDD in laptop PCs. Some storage manufacturers have also introduced flash drives in their low cost or midrange storage systems and are planning to introduce flash drives in their high end systems too. Even though flash storage still has some challenges in storage reliability to overcome it might become a viable solution to storage needs of archival community; its price per gigabyte is becoming competitive, it is more environmentally friendly due to lower demand for power, and it does not have moving parts, which could mean longer life time of storage units. A life time of ten years instead of five years for a storage unit could mean lower investment and management costs for an archivist since every other migration to the next storage technology could be skipped. In terms of read and write performance flash storage is already comparable with HDD technology.

6.3.22 Hierarchical Storage Management (HSM)

6.3.22.1 The OAIS Functions of Archival Storage embeds the notion of Hierarchical Storage Management (HSM) in the conceptual model. At the time OAIS was written the situation where large amounts of data could be affordably managed in other ways was not envisaged. The practical issue that underpins the need for HSM is the differing cost of storage media, e.g. where disc storage is expensive, but tape storage is much cheaper. In this situation HSM provides a virtual single store of information, while in reality the copies can be spread across a number of different carrier types according to use and access speeds.

6.3.22.2 However, the cost of hard disc has fallen at a greater rate than the cost of tape, to the point where there is an equivalency in price. Consequently the use of HSM becomes an implementation choice. Under these circumstances a storage system which contains all of the data on a hard disc array, all of which is also stored on a number of tapes, is a very affordable proposition, especially for digital storage systems up to 50 terabytes (and rising every year). For a smaller digital storage facility a fully functional HSM is consequently unnecessary and instead what is required is a much simpler system which manages and maintains copy location information, media age and versions and completely replicates the stored data on hard disc and on tape.

6.3.22.3 For medium to large digital storage systems the need for HSM storage systems remains and continues to be amongst the very expensive components of the digital storage systems.

6.3.23 File Management Software in smaller systems

6.3.23.1 The purpose of file management software in systems where the entire archive is replicated both on hard disc and tape is to keep track of the location, condition, accuracy and age of the tape copies. This basic backup functionality is a lower cost alternative to a classic HSM and may, at least in theory, be more reliable for small systems. However, as the large scale HSM represents a significant market, research and development has been supported by the industry in this area. Small scale file management software is being developed amongst the open source software development community. These include such systems as three most popular open source NAS applications, FreeNAS, Openfiler and NASLite, and the Advanced Maryland Automatic Network Disk Archiver (AMANDA). As with all such open source solutions, the onus is on the user to test the suitability and reliability of such systems, and without further development this publication makes no specific recommendation.

6.3.24 Verification and retrieval

6.3.24.1 In some commercial software, tape read/write error can be reported automatically during the data backup and verification process. This function is normally implemented with cyclic redundancy check, a technology using checksum against data to detect errors for transmission or storage. It is recommended that an error checking function should be implemented in any archival storage system. Error checking is difficult to implement in open source because that capability is linked to specific hardware. A commercially available stand-alone LTO Cartridge Memory Reader is the "Veritape" from MPTapes, Inc. and recently, Fuji Magnetics announced a Chip Reader Diagnostics System for LTO-Cassettes, bundled with software.

6.3.25 Integrity and Checksums

6.3.25.1 A checksum is a calculated value which is used to check that all stored, transmitted or replicated data is without error. The value is calculated according to an appropriate algorithm and transmitted or stored with the data. When the data is subsequently accessed, a new checksum is calculated and compared with the original, and if they match, then no error is indicated. Checksum algorithms come in many types and versions and are recommended, and standard, practice for the detection of accidental or intentional errors in archival files.

6.3.25.2 The cryptographic versions are the only type that have a proven record of trust when protecting against intentional damage to data, and even the simplest of these are now compromised. It has been recently shown that there are ways of creating meaningless bits that will calculate as a given MD5 checksum. This means that an external or internal intruder may replace digital content with meaningless data and that this attack will go unnoticed by the error checking management system until the files are required for use and opened. MD5, although still useful for transmission purposes, is 124 bit and should not be used where security is the issue. SHA-1 is another cryptographic algorithm that is under threat of being compromised, and which it has already been shown can, in theory, be circumvented. The length of SHA-1 is 160 bit: SHA-2 comes in versions with 224, 256, 384, and 512 bit lengths, and are algorithmically similar to SHA-1. The steady growth of computational power means that these checksums may, in the long run, be compromised as well.

6.3.25.3 Even with these compromises, a checksum is a valid approach to detecting accidental errors, and if incorporated into a trusted digital repository, may well be sufficient to uncover intentional damage to data files in low risk scenarios. However, where risks exists, and perhaps even where they do not, monitoring checksums and their viability must be part of preservation planning.

6.4 Digital Preservation Planning

6.4.1 Introduction

6.4.1.1 Once the action has been taken to convert the audio content to a suitable digital storage format for storage on a digital storage system, as defined earlier in this document, there is still a requirement to manage the ongoing preservation of the content. Section 6.3 Archival Storage includes a description of the issues surrounding management of the byte stream, i.e. ensuring that the digitally encoded data retains its logical structure through management of the storage technology.

6.4.1.2 There is, however, another aspect to the preservation of digital information, and that is ensuring that it is still possible to access the content encoded in those files. OAIS calls this function "preservation planning", and describes it as "the services and functions for monitoring the environment… and providing recommendations to ensure that the information stored… remains accessible to the Designated User Community over the long term, even if the original computing environment becomes obsolete" (OAIS 2002:4.2).

6.4.1.3 Preservation planning is the process of knowing the technical issues in the repository, identifying the future preservation direction (pathways), and determining when a preservation action, such as format migration, will need to be made.

6.4.2 Future Digital Pathways

6.4.2.1 When a file format becomes obsolete and is at risk of becoming inaccessible due to the unavailability of appropriate software to access the content, there are basically two approaches that can be made; migration, or emulation. In migration the file is modified, or migrated to a new format, so that the content can be recognised and accessed using the available software of the time. In emulation, the access or operating software is modified or designed so that it will open and play the obsolete audio file format on a new system which would not otherwise be able to open the content.

6.4.2.2 Our current understanding leads us to believe that for simple discrete files, such as uncompressed audio files, the most likely approach will be migration but this is not certain and all digital storage approaches and systems should be flexible enough to be responsive to the changing environment. Adequate preservation metadata as described in the PREMIS recommendations or the explicit file typing (including versioning) in BWF/AES31-2-2006 fields will support either approach, as will the standards being developed in AES-X098B which will be released by the Audio Engineering Society as AES57 "AES standard for audio metadata — audio object structures for preservation and restoration". Harvard University is developing a toolkit which supports the population of the necessary fields which will be released in open source.

6.4.2.3 This aspect of digital preservation is the strongest argument for an absolute adherence to the standard format described.The large investment the audio and IT industries have made in the standard audio format (.wav) means that the requirement for professional software tools which will enable the continued access to content will help to ensure that the sound archive can manage access to their collections. Likewise, the large investment in a single format will also help support the continuance of that format for the longest period, as the industry will not change an entrenched format without significant benefits.

6.4.3 Motivating Factors and Timing

6.4.3.1 Though the wise choice of standard formats, and an observance of industry practices will delay the eventuality, the day will come where it will be necessary to undertake a preservation action of some type which will be needed to maintain access to the audio content stored. The issue for sound archivists concerned with their digital content will be determining when to undertake that step and what precisely to do.

6.4.3.2 A number of initiatives are being developed to help support this need. These include the Global Digital Format Registry (GDFR http://hul.harvard.edu/gdfr/), which exists to support "the effective use, interchange, and preservation of all digitally-encoded content." Other services provide recommendations about suitable format, such as those provided by the Library of Congress (US) or The National Archives (UK).

6.4.3.3 The factors which will motivate a sound archivist to undertake some sort of preservation action will be the recognition that new software no longer supports the old format, and the industry as a whole moving to select a new format. Knowledge of the events that herald change comes from expert understanding of the technology, the industry and the market and sound archivists are well advised to take heed of the recommendations services such as those noted above.

6.4.3.4 Software and services under development, such as the Automatic Obsolescence Notification System (AONS), will provide advice to collection managers on when changes have occurred in the market requiring action (https://wiki.nla.gov.au/display/APSR/AONS+II+Documentation). The implementation of such services will occur in parallel with the development of the GDFR.

6.5 Data Management and Administration

6.5.1.1 Data Management, in the OAIS, is the services and functions for populating, maintaining, and accessing both descriptive information which identifies and documents archive holdings and administrative data used to manage the archive, in other words the catalogue of content and the statistical record of data content.

6.5.1.2 Administration, in the OAIS, is the services and functions for managing system configuration, monitoring operation, providing customer service and updating archival information. It is also responsible for management processes such as negotiating submission agreement with producer, auditing submission, control physical access, establishing and maintaining archive standards.

6.5.1.3 The management and administration of the digital repository and archival system provides services that allow the sustainability of the system and the preservation of the content stored therein. A requirement of an archival digital storage system would include the ability to interrogate the system to produce result sets of holdings, access usage statistics, contents summaries including sizes and other necessary technical and management information. The data management and administration is critical to a sustainable archival system because this functionality ensures that files preserved and accessed are properly found and identified.

6.5.1.4 It is within this section of the digital storage and preservation system that control over access to content, or security control, is implemented. Many repository software systems incorporate approaches to implementing policies which are stored and managed by the system. It is important to recognise that the rights management information, like the audio content itself, must outlast the system used to store it, and so be transferable to any future replacement preservation and storage system. Information which is encoded in XACML (*eXtensible Access Control Markup Language*) for example, is both more universally enforceable, and transferable to other systems. XACML is a declarative access control policy language implemented in XML and a processing model, describing how to interpret the policies. XACML is managed by the OASIS standards group (http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml).

6.5.1.5 When selecting, establishing and installing a digital preservation system one of the critical tests should be to determine if the administration of the proposed system is within the capabilities of the host institution. The capability and breadth of functions of a system is often linked with the complexity of use and installation. A system which cannot be adequately managed and maintained is a major risk to the content it manages. It is therefore important that the long term management of a system take account of the available technical expertise required to sustain its use.

6.6 Access

6.6.1 Introduction

6.6.1.1 The OAIS Reference Model defines "access" as the entity that "provides the services and functions that support consumers in determining the existence, description, location and availability of information stored in the OAIS, and allowing consumers to request and receive information products." In other words, access is the mechanisms and process where content is found and retrieved. IASA-TC 03 "The Safeguarding of the Audio Heritage: Ethics, Principles and Preservation Strategy" makes the point that "the primary aim of an archive is to ensure sustained access to stored information". The preservation of the content is a prerequisite to sustained access to the content, and in a well planned archive access is a direct outcome of it.

6.6.1.2 In its simplest form, access is the ability to locate content and, in response to an authorised request, allow retrieval of the content for listening, or possibly, as long as the rights associated with a work allow it, creating a copy that can be taken away. In the connected digital environment access can be provided remotely. Access, however, is more than just the ability to deliver an item. Most technically constructed archival systems can deliver an audio file on request, but a true access system provides finding and searching capability, delivery mechanisms and allows interaction and negotiation regarding content. It adds a new dimension to access beyond that of conquering distance. In this new services based model of retrieval, access could be considered a dialogue between the provider's system and the user's browser.

6.6.2 Integrity in On Line and Off Line Access Environments

6.6.2.1 Prior to the existence of remote access in the online environment, such things as authenticity and integrity were established by individuals in the reading rooms and listening posts of the collecting institutions. The content was delivered by representatives of institutions whose reputation spoke for the integrity of the content. Original materials could be retrieved for examination if the copies were questioned.

6.6.2.2 The online environment still relies to some extent on the trusted nature of the collecting institution, however, an unambiguously original item can never be provided online, and the possibility of undetected tampering or accidental corruption exists within the archive and distribution network. To counter this, various systems exist which mathematically attest to the authenticity or integrity of an item or work.

6.6.2.3 Authenticity is a concern with knowing that something has originated from a particular source. The trusted nature of the institution creating the content attests to the processes, and a certificate authority is issued which a third party can use as a guarantee of authenticity. Various systems exist and are valuable where this could be an issue.

6.6.2.4 Integrity refers to a wish to know whether an item has been damaged or tampered with. Checksums represent the common way of dealing with integrity, and are valuable tools in both the archive and the distribution network (see 6.3.23 Integrity and Check sums). However, as is discussed in 6.3.23, checksums are fallible, and their use requires monitoring on behalf of the archive of latest developments.

6.6.3 Standards and Descriptive Metadata

6.6.3.1 Detailed, appropriate, organised metadata is the key to broad exposure and effective access. In Chapter 3 Metadata, a detailed discussion of metadata in many of its forms and requirements is undertaken, and this should be referred to in developing a delivery system. Ambitious access facilities, using, for example map interfaces or timelines, will only function if there is metadata to support it in a structured and organised form.

6.6.3.2 The most cost effective way to manage and create the appropriate metadata is to ensure the requirements for all the components in the delivery system are established prior to the ingest of the content. In this way the metadata creation steps can be built into the pre-ingest and ingest workflows. The cost of creating a minimal set, as discussed in Section 7.4, is the extra task of adding and structuring the metadata in a system which has already been created.

6.6.4 Formats and Dissemination Information Packages (DIP)

6.6.4.1 The Dissemination Information Package (DIP) is the Information Package received by the Consumer in response to a request for content, or an order. The delivery system should also be able to deliver a result set or a report from a query.

6.6.4.2 Web developers and the access "industry" have developed delivery systems based, naturally, around delivery formats. Delivery formats are not suitable for preservation, and generally, preservation formats are not suitable for delivery. In order to facilitate delivery, separate access copies are created, either routinely, or "on demand" in response to a request. Content may be streamed, or downloaded in compressed delivery formats. The quality of the delivery format is generally proportional to its bandwidth requirements, and collection managers must make decisions about the type of delivery formats based on the user requirements and the infrastructure to support delivery. QuickTime and Real Media formats have proven to be popular streaming formats and MP3 (MPEG 1 Layer 3) a popular downloadable format which may also be streamed. There is no requirement to select only these formats for delivery, and many collection delivery systems provide a choice of formats to the user.

6.6.4.3 For some types of material it may be necessary to create two master WAV files: one, a preservation or archival master that replicates exactly the format and condition of the original the second, a dissemination master that may have been processed in order to improve the audio quality of the content. A second master will allow the creation of dissemination copy as required. It is expected that distribution formats will continue to change and evolve at a faster rate than master formats.

6.6.5 Search Systems and Data Exchange

6.6.5.1 The extent to which content can be discovered sets the limit on the amount of use of the material. In order to ensure broad usage it is necessary to expose content through various means.

6.6.5.2 Remote databases can be searched using Z39.50, a client-server protocol for searching and retrieving information. Z39.50 is widely used in the Library and Higher education sector, and its existence predates the web. Given the extent of its use, it is advisable to establish a Z39.50 compliant client server on databases. However, this protocol is being rapidly replaced in the web environment by SRU/SRW (Search/Retrieval via a URL and Search/Retrieve Web service respectively). SRU is a standard XML-focused search protocol for Internet search queries, utilizing CQL (Contextual Query Language), a standard syntax for representing queries (http://www.loc.gov/standards/sru/). SRW is a web service that provides a SOAP interface for queries in partnership with SRU. Various open source projects support SRU/SRW in relation to the major open source repository software such as DSPACE and FEDORA.

6.6.5.3 OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) is a mechanism for repository interoperability. Repositories expose structured metadata via OAI-PMH which is aggregated and used to support queries on the content. OAI-PMH nodes can be incorporated into the common repositories. OAI-ORE (Object Reuse and Exchange) will be important for the sound and audiovisual archiving community as it addresses the very important requirement to be able to deal efficiently with compound information objects in synchronisation with Web architecture. It allows the description and exchange of aggregations of Web resources. "These aggregations, sometimes called compound digital objects, may combine distributed resources with multiple media types including text, images, data, and video". http://www.openarchives.org/

6.6.5.4 In order for the sophisticated online environment to work it is necessary to have interoperable metadata and content. This means that there must be some shared understanding of the attributes included, a general schema which is able to operate in a variety of frameworks, and a set of protocols about exchanging content. This is best achieved, as is always in the digital environment, by adhering to the standards, schemas, frameworks and protocols recommended and avoiding proprietary solutions.

6.6.6 Rights and Permissions

6.6.6.1 It is important to note that all access is subject to the rights established in the items and the permission of the owner to use the content. Various rights management approaches exist, from "fingerprinting" the content, to managing the permissions of various individual to access, the physical separation of the storage environment. The particular implementation rights system will depend on the type of content, the technical infrastructure and the owner and user community and it is beyond the scope of this document to define or describe a particular approach.