Deliverable   D13.4 Repositories

Digital Repositories Explained

DOCUMENT IDENTIFIER	PS_WP13_BBC_D13.4_Repositories_v1.0
DATE	27/2/2007
ABSTRACT	This document sets out the requirements for a secure, sustainable *digital repository* – with special reference to digital audiovisual materials – and examines the technology and standards being developed to fulfil those requirements. It recommends "OAIS for datatape" as the needed (and missing) approach.
KEYWORDS	audiovisual digital repository trusted requirements standards technology preservation
WORKPACKAGE / TASK	WP13 - Preservation and Access Planning
AUTHOR, COMPANY	Richard Wright, BBC
INTERNAL REVIEWERS	Adam Lee; Steve Jupe; Ant Miller, Jean-Hugues Chenot

DOCUMENT HISTORY

Release	Date	Reason of change	Status	Distribution
0.9	10/01/2007	First Complete Delivery	Living	Restricted
1.0	22/02/2007	Finalised, ready for publication	Complete	Public

Contents Table

1. Document Scope and Summary

There is a cluster of institutions that share a common problem: their whole method of operation will be changed by the move from conventional media to all-electronic media. All institutions which maintain long-term collections of paper, recorded sound and recorded images have to make the shift to managing long-term electronic storage. The core issue is defining – and meeting – the requirements for a secure, sustainable *digital repository*.

This document sets out those requirements, and examines the technology and standards being developed to fulfil the requirements. Particular attention is given to audiovisual archives and their associated (or planned, or wished-for) digital repositories.

In the section on costs (Section 8) we discover that repositories, at least as so far implemented in the libraries world, are too expensive for use as a complete replacement for conventional 'tapes on shelves' collections. PrestoSpace recommends migration to the 'cheapest acceptable digital storage media'. For the next decade at least, the choice will be discrete media, mainly data tape, on shelves or in robots. However the content written on the media will be files, and files need "repository functionality" in order to remain understandable (playable) indefinitely. The essential functionality is to acquire, store and deliver 'information packages' rather than just files. The extra information provides **the life-support system** for the file – the information necessary to understand the data in the file, and to migrate it as required to future file formats. The standards (mainly OAIS) defining the required functionality are presented in Section 6.

The need is: a way to implement OAIS functionality in the absence of mass storage (servers; everything on spinning discs): meaning what is needed is **OAIS for data tapes** on shelves (or in robots) – which so far has never been explicitly implemented. The proposed *PrestoSpace Competence Centre*[1] would be essential for changing analogue collections into data tape collections using OAIS processes, because the Competence Centre would specify the 'information packages' for audiovisual material, supply information on audiovisual formats, recommend new formats, provide the information required to move from old to new formats – and finally supply the guidance and training archives would need in order to us 'information package' functionality.

2.   Definition of Digital Repository

A repository should be a safe place to keep things, and for a digital repository, those "things" are digital objects: files. But definitions also include:

- "An organisation that has responsibility for the long-term maintenance of digital resources, as well as for making them available to communities agreed on by the depositor and the repository" Research Libraries Group[2]

- "A collection of digital assets and/or metadata accessible via a network without prior knowledge of the digital repository's structure. … A repository is managed by a data provider to expose metadata to harvesters." [3]

So a digital repository is:
1. a **place** (to put things)
2. an **organisation** (to set up and maintain things)
3. a **collection** (of assets; the things in the place)
4. a server and network or other **mechanism providing electronic access** (to the things)

The multiple levels or aspects of a digital repository matter, because they are all part of the essential purpose: a safe place for digital objects, a safe place to keep and use digital objects, a safe place to go back to (years, decades and we hope centuries later) and still find usable digital objects.

The following sections look at these various aspects of *digital repository* in more detail, including formal definitions and relevant standardisation.

2.1.   A place (to put things)

A digital repository is a physical thing – it holds files. The location of the storage may not be obvious – and the storage may be in several places and several technologies – but still there must be storage somewhere. In addition to storage devices (hard drives, data tape robots and other equipment), a repository has functions (operations ; services) and rules for the operation of those functions.

Storage without rules would be like a library with shelves and nothing else:
- no process for acquiring books
- no way to get data about the books into a catalogue
- no catalogue!
- no control over how books are taken off the shelves

Acquisition, cataloguing and circulation functions are essential to any library – and the same principles apply to a digital repository. Raw storage is like raw shelves, and it takes a whole structure of rules and services on top of the storage in order to have a useful repository. The conclusion is that the storage is the least significant

part of a digital repository: it can be anywhere, on any medium – so long as the storage doesn't fail, and supports the overall rules. Detailed storage requirements are set out in Section 5.

Libraries have lots of rules about cataloguing and other basic processes, set up over centuries[4]. Digital repositories are very new, but there is one major standard covering the processes for a digital repository: OAIS (presented in detail in Section 6).

2.2.   An organisation (to set up and maintain things)

The rules and services require an organisation – to set up and maintain the operation. The rules and services could be fully automated, so a digital repository would not imply staff or a building – but somebody has to implement the repository: commission it in the first place, maintain it forever after.

In conclusion: the 'place' aspect of a digital repository is not the storage location, as that could be anywhere, and could be distributed. **The essential 'place' is the institution in charge.** The institution could be manifested just as a web activity (no building, no permanent staff), but there must be some "permanent controlling entity" to ensure continuity of services.

There is general recognition of the importance of the institution behind a repository. Secure storage and access over the long term requires a stable institution. This requirement is recognised by the professional bodies developing standards and 'best practice recommendations' for digital repositories. In 2000-2001, the US-based Research Libraries Group set out the general 'attributes' supporting trust, in their document *"Trusted Digital Repositories: Attributes and Responsibilities"*[5]. Later work has developed a detailed checklist for assessing these attributes: *"Audit Checklist for Certifying Digital Repositories".*[6]

2.3.   A collection (of assets; the things in the place)

A collection of assets is more than just the assets themselves. For a pile of items to become a collection requires documentation and management of those items. The whole profession of archiving – and the allied professional skills of librarianship – are relevant to the distinction between a pile of items and a collection.

In particular, a collection will have documentation – which could be at two major levels.
- item level – a catalogue of the contents of the collection.
- collection level – an overall description of the collection.

The requirement for documentation and management is just as valid for electronic collections as for non-electronic, and the standards and best practices are also equally valid.

A principal difference between electronic and non-electronic collections is in the importance of the documentation and management. A collection of tangible items in one place, like 700 Greek vases in a storeroom of a museum, can be expected to exist indefinitely irrespective of documentation. Much of its potential value will disappear if the provenance and original location and general history of the items is lost – but much of their value will persist.

For electronic files, virtually all their value could be lost if the documentation is lost. It could be literally impossible to identify voices or images, so a bunch of audiovisual files on a forgotten hard drive will have very little 'artefact' value, if any. Hence documentation and management – the 'value added' that defines a collection – are vital. *Any digital repository worthy of the name has documentation and management as essential functionality, ensuring that the collection is preserved, not just the items.*

In summary, the difference between storage (as understood by the IT industry) and a true *digital repository* is the difference between a pile of items and a true collection.

2.4.   A method of access (to the things)

In addition to the above definitions, a repository is a place to go to get things. When a repository is a building, a place to put things (definition 1) is the same as a place to get things. With electronic access, the 'putting' and 'getting' part company. Storage is an invisible service, and electronic access is through something like a website or portal.

There are two steps to the access:
- **Access to the "place"** – to the electronic repository overall, which involves matters like knowing the URL, and finding one website amongst millions. Key technology includes general *search engines, portals* and equivalents of *union catalogues* – all the technology that takes place **above** the level of the individual website.
- **Access to the "things in the place"** – once the overall collection is found, as a website or equivalent, any sizeable collection then requires a search of the collection to find desired items. This is the access level for all technology that operates **at or below** the collection level. For a repository, this technology will include the online catalogue and search-and-retrieval technology built into the repository.

Electronic access has created a new problem for archives and other collections: is it the content that matters most, or the institution that maintains the collection and provides the content?

The question arises because electronic access can be organised either way. With tangible items in a building somewhere, the issue never arose.   The user had to go to the institution first, to then get access to the items.

With electronic access the same two stages can be involved – or there is the new possibility of allowing the users to go directly to the items, without ever being aware of the institution and collection that makes it all possible. As an example, a person seeking instances of American ragtime music on wax cylinder recordings could first

go to major collections of wax cylinders digitised and available online – such as the collection of the library of the University of Santa Barbara, which is here: http://cylinders.library.ucsb.edu/

As a second process, the online catalogue at that website can be used to find 91 recordings indexed as 'ragtime' – and ten as 'ragtime' and 'berlin' – meaning Irving Berlin (1888-1989) whose career as a composer spanned 8 decades from wax cylinders to CDs.

Alternatively, using the general web search engine Google there are 462 hits for the search: **irving berlin wax cylinder ragtime** – and there are 157 for the search: **irving berlin wax cylinder ragtime mp3**.   The second search should focus on websites that supply a downloadable mp3 file. In both cases the first hit was the same: *Internet Archive* which has online content from various American historical collections provided by the US National Park Service. Hits 8 and 9 on the Google search click-through to the University of Santa Barbara collection – but straight to search results rather than to the top level (home page) of the site.

There are two points from the above example that should matter to collection owners:
- a user of a general search engine may go to somebody else's collection to find particular content
- a user of a general search engine may come directly to a 'search results' page in your collection, bypassing all the usual introductory and search interface pages. This route can leave the user quite unaware of the identity of the collection, because the user basically does a 'hit and run' – grabs the desired content and exits.

The issue of collection vs content and how best to use the web is very large, and applies to everything on the web that is accessed through localised search functionality such as a collection catalogue. It certainly isn't an issue just for repositories.

Repository managers thus have a whole series of issues regarding web access:
- should they hide content from search engines, to prevent the 'bypass the home page' phenomenon just described?[7] [8]
- should they actively promote supply of content description to search engines, to encourage the greatest amount of access, regardless of whether or not users understand exactly where they are getting content?[9] [10]
- should they work with other institutions to share content descriptions, forming a **portal**[11] (common access website) or **union catalogue?**[12] [13] [14]

All of these are considerations at the repository level – concerning how to survive and thrive on the web. They are managerial issues that are outside the scope of what is meant to be an explanation of digital repositories, and so will not be discussed further. The references just given are a starting point for further information.

3. Electronic Access

A main reason for having a digital repository is to provide an improved alternative to media on shelves, in order to take advantage of electronic methods of access.

Clearly electronic access offers obvious improvements:
- speed of delivery
- elimination of contention for copies
- elimination of circulation control: booking, tracking, overdues

Not so obvious are the new methods for access, and new reasons for access. These are not so obvious because they may not yet exist, or be at an early, primitive stage.

There are two main parts of digital access:
- web access
- electronic delivery

3.1.   Web Access

Access to a digital repository is electronic – a data connection between a user and the repository. It can use various methods, but the main one is web technology.

The web needs highlighting because of several factors. The web:
- is in people's homes (in Europe)
- has come to define what does and does not exist
- has spawned a variety of search mechanisms
- is becoming the preferred source for at least some audiovisual material, beginning with commercial music tracks.

The combined result of these factors is as follows: audiovisual collections will need to be 'on the web' to show that they exist; users from across the world will find them; these users will expect access to their contents.

An audiovisual collection thus is in a difficult position: if not on the web, it effectively doesn't exist. If on the web, it raises an expectation that, largely, it cannot fulfil.

Meeting the *expectations of the web* is likely to be the principal challenge of the next decade, for audiovisual collections. It follows that providing for web access, in terms of 'discovery metadata', the organisation, quality and standardisation of that metadata, and in terms of delivery of actual content at various qualities and bandwidths, will be principal requirements of the repository.

Some repositories may not welcome public access. The national mint doesn't open its doors, and some digital repositories (for instance, for legal, financial or medical records) may need to implement restricted access to their digital contents. There are alternatives to the web, and there are special cases of restricted (rather than public) use of web technology.

The technologies for private electronic access[15], or restricted use of web technology[16], are again a general topic not in any way restricted to digital repositories. No further discussion will be presented here, but the cited references can be consulted for more information.

3.2.   Electronic delivery

Audiovisual archives have been notoriously inaccessible, because the contents of the archive were on professional audiovisual formats that required careful handing and expensive equipment for playback. Electronic access to audiovisual content solves a number of problems simultaneously:

- **Contention** for copies: elimination of the 'somebody already has it' problem, and elimination of the need to take steps to avoid that problem, such as making multiple viewing copies

- **Circulation** of control: checking out, delivering, and ensuring the return of physical media. However elimination of physical circulation could result in elimination of all knowledge of who has accessed material, so a new sort of circulation control is a requirement of a digital repository

- **Physical attendance** at the archive: electronic delivery can go anywhere (given adequate bandwidth). However if people are no longer presenting themselves within the walls of an audiovisual collection, the new problem is introduced of how to know who is entitled to access the material. Indeed, the whole restriction of access to 'legitimate research', which is built into the rules and laws of many collections, is called into question by electronic delivery.

- **Damage** to the material. Essentially electronic delivery makes 'viewing copies on demand', so the 'original' – however that is to be defined in a digital environment – is unaffected by access. This statement needs qualification, because access to digital files on optical or magnetic media does potentially involve wear – although in practice it may prove far better to read files, and check error rates, than to leave files unread an not know about physical deterioration within the repository. Indeed digital repositories may be no more immune from media wear than conventional tapes on shelves – it is just that they should have automated, efficient, cheap and reliable methods for checking and regenerating data.

In short, electronic delivery removes *all* the physical, technical and logistical barriers to *unlimited* access to audiovisual collections. The effect will be to make the legal barriers all the more obvious – they will be all that's left. Inevitably, all the pressure to hear and see audiovisual material – from a world audience – will be applied specifically to these legal barriers. Given the amount of change to the physical barriers in just the last few years, it would be reasonable to expect that the next one or two decades will see significant changes to the legal situation. The implication for repositories is that their electronic delivery capabilities may have to grow far beyond the current access requirements for today's audiovisual archives.

3.3.   Search and Retrieval

Strange as it may seem, the ability to find things is not core functionality of current thinking about digital repositories, at least at the 'functionality standardisation' level of OAIS (see section 6.1). The standardisation activity around digital repositories has concentrated on preservation, not on access.

In practice, any working audiovisual repository will inevitably have two parts:
- the *digital repository* itself, which is a kind of bullet-proof, bolt-down, ultra secure approach to permanence of stored items
- *something else,* which is optimised for web access to browse-quality audio and video

The *something else* will likely be a separate copy of the descriptive data in the repository, in a system designed for effective search and retrieval. The *something else* will also include separate copies of the repository content, but in the from of lower-resolution *proxies* – certainly NOT the original files and not even exact copies of the originals (because the originals will presumably be at highest available quality, and the proxies will be whatever quality and encoding is currently desired for web access).

Once this two-part approach is established, search and retrieval becomes a separate issue from the establishment and operation of the digital repository. Of course they must stay in synchrony, but the digital repository is all about robust and fool-proof processed for putting material in and taking it out – so all that is needed is ensuring that these 'in and out' processes include update of the *something else* that sits outside the repository.

Audiovisual search and retrieval is therefore not considered further here – but it is a main concern of PrestoSpace. Relevant PrestoSpace work includes the following:

- D15-3 Content Analysis Tools for Video, Audio and Speech Report[17]
- Deliverable 15-1 Analysis_AV_documentation_models Report[18]

The leading international work in audiovisual search and retrieval is associated with the annual research evaluation exercise TRECVID, which brings together 100 or more research activities, covering both the university and the commercial sectors. Further information about TRECVID is here: http://www-nlpir.nist.gov/projects/trecvid/

4.   Why Digital Repositories Matter

Digital repositories are important because:
1. digital objects are more 'at risk' than non-digital objects
2. many interesting and useful physical objects are being converted to digital ones (books, journals, photographs, audiovisual recordings)
3. many new interesting and useful objects are 'born digital'

4.1.   Digital objects are more 'at risk'

There are some terminology issues about what is a 'digital object' – because CDs, DAT audio tapes and digital video tapes (such as a Digibeta) are certainly physical objects. These 'hold in your hand' forms of digital media have there own preservation problems, of deterioration and particularly of device (format) obsolescence. But analogue audio and videotape has related problems, and in any case a digital repository is not a solution.

The real issue is around *files,* the basic unit on digital storage. There are various new risks that apply only to digital files:

- A file is not a tangible object, and that is one new source of risk. Intangible (electronic) objects can be instantly erased from their carriers. So-called permanent storage or write-once-read-many storage attempts to minimise the risk of erasure).

- The other principal new risk with files is that they can be easily lost. The file itself cannot ever be found directly – the bits themselves need to be identified as an entity in some sort of file management system. If that system fails, or if any of its component tables or directories fail, the file is irretrievably lost even if the bits still exist, unaltered, somewhere in the storage system.

- The file may be fine, and the file management may be fine, but at a higher layer the access technology may not understand how to process the file. An 'unknown file type' or 'unrecognised format' or some such error may appear. Such a problem may have a solution, but it could take much effort to provide the right software to 'play' (interpret) the file.

- The above are all clear risks for audiovisual media, and for all files. For more specialist media such as databases and computer programmes, there are additional risks around losing the high-level information required to interpret or run the data. As audiovisual files become more complex, as through wrappers such as MXF, such interpretation risk – that go well beyond an issue of finding an appropriate player – will only increase.

The good news is that audiovisual files share the digital preservation problems of all types of file, and so there is a whole industry developing around digital preservation. Further information is available from the Digital Preservation Coalition[19] in the UK,

the Library of Congress[20] [21] in the USA and from the PADI[22] centre in Australia, the ERPANET[23] and DPE (Digital Preservation Europe)[24] projects, amongst many others.

4.2.   Lots of material is digitised

Most institutions which are in any way a library or archive are involved in digitisation projects[25]. The process began with scanning of documents and photographs in the 1980's, and blossomed with the development of the internet. With the rise in broadband internet access, audiovisual collections can now use the same technology to deliver high-quality audio and reasonable-quality video straight into private homes – as well as educational and commercial institutions.

The processes and technology of digitisation arose as an alternative to microfilm and microfiche. Microfilm[26] was widely used in document and photograph (and general 'flat media' collections) for making a second version (proxy) used both for preservation and for access.

With the rise of computers on desktops, and networks to link those computers, large institutions (eg research organisations, university libraries) began to use digital proxies because they had advantages in terms of access: electronic distribution across the network. Electronic copies could also be duplicated at virtually no cost.

With the internet, these advantages multiplied as "the network" expanded to cover more and more of the world.

In the last decade, digitisation has been accelerated by a new kind of institution: the new-technology information intermediators and aggregators. Aggregators of images include Corbis[27] and Getty Images[28], for audio the main company is iTunes[29] (Apple), and for video there are various activities emerging – ranging from parts of Corbis and Getty that now deal with audiovisual materials to user-generated collections – the most famous being YouTube[30].

Digitisation and aggregation have also been undertaken on a large scale by non-commercial projects, including The Million Books Project[31] and the work of the Open Content Alliance[32] and the Internet Archive[33]. Among the many interesting aspects of the work of the Internet Archive – which should be known to all because of its enormous significance – is the special permission they acquired from the US Congress to digitise and make public "orphan" books -- out of print but not out of copyright, and therefore lost as regards public access except for the innovative work of Internet Archive. The Internet Archive now has 100,000 books online, as well as 85 billion internet pages and much more.

On the commercial side, Google Books is working on digitisation of the complete contents of six of the world's major libraries, making the text searchable – though not the complete contents except for out-of-copyright books.

4.3.   Most media is now 'born digital'

The change to digital technology is no longer for the future. In large part most new materials in all media are now produced digitally, even if their final distribution may be in non-digital media: films, newspapers, books.

- Text: personal as well as business documents (including most newspapers and books) and all email
- Images: many types of film are now discontinued, and most new cameras are digital
- Audio: CDs replaced tape starting 20 years ago. The music industry has been using digital technology for so long and so extensively, that CDs are rapidly being replaced by direct download of audio files, such as MP3 – and the end of the CD has been predicted (by one source) by 2012[34]. Telephones converted from analogue to digital coding and distribution starting in the 1970's
- Video: the last main consumer analogue format, VHS, is obsolescent – the machines are no longer made. At the professional level, digital videotape began to take over from analogue in the 1990's. BetaSP is the last Sony analogue format and will be made completely obsolete as television production switches to high definition[35]. Indeed broadcasting is moving away from tape altogether, and "digital production" is becoming synonymous with "server-based", "network-based", "tapeless" and the really unfortunate term "non-linear" as jargon words for using computer-based methods and storage rather than using videotape and videotape recorders.
- Film: "The Death of Film"[36], like that of Mark Twain[37], may be exaggerated – but film production is already largely digital, going to film only for the final stage: distribution. One reason is that special effects require digital technology, but even standard editing and other production techniques, such as colour management, are becoming cheaper and better with digital technology. Film is thus following the same route as newspaper and book production: produced digitally, distribute non-digitally.

New media is digital: internet web pages, optical discs of various sorts, datatape, memory cards, digital audio broadcasting, digital TV broadcasting, mp3 download and players, mobile telephones, 'voice over internet', Bluetooth, wireless interconnectivity – these are all digital. There is a continuing interest in old media, such as vinyl recordings, but there is virtually no production of new releases on old media (apart from film) – and there is almost no development of new forms of analogue media. Analogue television is being 'switched off'. Film will eventually – at least it part – convert to digital display technology (digital cinema), at with point distribution will be by files, not on film. There is a strong possibility that analogue radio will continue indefinitely – but all means of production of analogue radio, and of recording it for archiving and later access, will be digital.

5. Requirements for an Audiovisual Digital Repository

What archives want from storage is a question that cannot be answered in isolation because the fundamental question is about what the archive needs and wants to do. So there is a range of requirements, at multiple levels:

- The requirements of archives start with the **archive service requirements**: what an archive is good for; what an archive does.
- Below that are the **functional storage requirements**: the *function* the storage fulfils. Successful digital archives will use storage that functions adequately, storage that serves its purpose.
- Below that are **storage service requirements** – The technical requirements about gigabytes and bandwidth.
- Finally there are **storage media requirements**: how storage operates. The archive has to have some physical reality, somewhere. The contention of this paper is that archivists (well, some of them) will initially be interested in digital media, but will quite quickly move back up to being primarily interested in the service, not the media.

Archives perform services – or they're of no use and risk disappearing. As archives move from a *storehouse* to a *service provider* perspective, they move away from storage as a primary activity. Ultimately, archives and storage devices will part company. Digital archives will use a *storage service provider*, just as so many other IT functions now use service providers of one sort or another (ranging from networks to data centres). But digital archives need a service provider which understands archives and understands

Figure 1 - Mass Storage, 1960's style

storage – particularly long-term storage – and it is surprisingly hard to get such expertise from the standard IT industry.

5.1. Storage Requirements

Regarding how storage works, archives really only want information in two areas:

- **Persistence:** the ability to get content out of storage;
- **Currency:** the ability to use that content.

These two terms are not standard in the storage industry, but they are basic concepts (under various labels) of digital preservation technology – and from the work on storage done under the SAM (Storage and Archive Management) part of EC project PrestoSpace[38]. I use these concepts rather than standard storage

terminology, because of the mismatch between the information archives want, and the statistics generally available.

5.1.1.   Persistence

Persistence is not a standard term in either archives or in the storage industry, although it is a standard IT term in the context of the Worldwide Web:

- **Persistent** identifiers[39]: the various efforts within web technology, to counter the general tendency for resources on the web to 'go missing'(broken link; dead link; link rot; 404 error). Estimates vary, but figures around 30% (almost regardless of the context) are common[40].
- **Persistent** resources[41]: the links or identifiers are a means to an end; persistence of the resources is the real issue.

Archives have exactly the same two concerns: not losing their metadata, and not losing the archive contents. The metadata is just a way to get to the content – identifiers, descriptors and finding aids.

Digital archives will have files to hold the content, and a storage system using sort of addressing scheme to locate the content. Thus, digital archives share the problems of the World Wide Web and the IT industry, specifically the concern for persistence.

The storage industry does not use the term *persistence*. Typically, storage industry information relevant to losing stored data is expressed in terms of error rates (of the data reading process), failure rates (at the device level) and media life expectancy.

There is a real gap here, because archivists have NO interest in read error rates and MTBF, and it is a conjecture of this paper that 'digital archivists' will also have no interest in media life expectancy. Meanwhile the storage industry provides data about storage media systems, and NOT directly about the persistence of the content.

The author's view is that the storage industry provides this sort information because that is the easy thing to do. They have information on media and systems, on its performance and failures. When the storage industry talks to archives, they should consider providing the information archives really want: will the content still be there in 20 years? Will it persist?

Specifically, archivists want to know:

- **How much content will be lost**, every year for N years? This is the one figure an archive can use to decide whether or not a storage strategy is acceptable.
- **What is the statistical distribution of the probability of loss**? This information allows an archive to assess the degree to which performance (of the storage strategy) can be trusted. It's no good investing in a strategy with a 1% projected loss, if there is a 50% chance that the loss can be 10 times higher. This may look a bit complex and exotic – statistics about statistics – but it's exactly the same complexity of information an insurance company uses to compute life assurance premiums. Only when an archive knows the confidence interval around the probability of loss can it make informed decisions about control of risk.

- **How do the probabilities vary with N (how do they vary over time)?** This is again basic information, because storage strategies need to be re-assessed regularly. It may well make sense to change strategy after a shorter rather than a longer time, because probability of loss may well increase over time (or the costs – of keeping losses from rising – may themselves rise). A good horse to bet on can, in time, turn into a tired horse or an expensive horse. We are all familiar with this situation, especially with respect to being a car owner. Consumer guides to car ownership provide relevant information. Archives would like the same sort of information from storage providers.
- **How do the probabilities vary with cost?** We all expect to 'get what we pay for'. We fully expect that a storage strategy with 99% persistence over 20 years would cost more than one with 95% persistence. How much more? Archives simply cannot get this information – not because the vendors won't say, but because the storage industry simply does not compute the statistics that the archivists most want to see.

5.1.2.   Currency

Currency is also not a standard term, as terminology in the digital preservation area is still being established. The problem is format obsolescence, and currency refers to whether a storage strategy can deliver data *usable* content – usable by current technology.

There is much work in digital preservation on format obsolescence. It is a recognised problem, and much has been done to develop and implement solutions. For digital files in general, major institutions such as The National Archive in the UK and the US Library of Congress are developing software repositories[42] for legacy software. Many institutions are developing strategies to keep content usable (eg UKOLN in the UK[43], PADI[44] in Australia).

Persistence is a dimension of digital archive storage where archive can expect the storage industry to come up with relevant statistics. The issue of *currency* is more difficult, but the whole digital library and digital preservation community has identified this problem and is working on solutions.

A digital broadcast archive has two main choices:
- keep the original content as is, and ensure that there will always be players available to render the content into usable audio and video signals;
- migrate the content as formats become obsolete.

The first option is fraught with problems, as it is an immense ambition to make players not only available, but to have those players where they are needed, namely alive and working on the desks of the archive users.

The second option is a chore, but one that keeps content viable. It the migration route is chosen, persistence is also affected – because updating files for currency requires that the files be read and re-written, which is a basic 'refresh' operation that could well be a cornerstone of the strategy for persistence.

Ideally, the storage industry would supply information covering costs of such 'refresh' operations, so that an archive could balance the benefits (for both persistence and currency) against costs of such a major operation as re-formatting an entire audiovisual collection. In practice, the storage industry does not supply this information – because it is about the use of devices rather than about the devices themselves, and so it is 'the customer's business'.

The currency issue takes precedence over the simple statistics that the storage industry does provide. What is the advantage (for anyone) of a medium that will hold a file for 100 years, or even 40 years, if the file format itself becomes unplayable within 10 years? As a reminder of the problem, what proportion of document or PowerPoint files from 1996 can be opened today? If it is less than 99.5%, then it is below the minimum persistence level likely to be required by archives.

5.2.   Storage Service Requirements

Storage may seem to be the central issue to a digital archive, but persistence and currency are the essentials. The remaining technical requirements are only two:

- Size of the storage
- Bandwidth of the access to the storage

Size is the easier of the two. The complications are decisions about file formats and degree of compression to be used on master-quality files (if any – archivists hate compression!). A "ready-reckoner" for calculating storage requirements is available on the PrestoSpace SAM website[45].

Storage size and file *persistence* are related. The more copies, the more redundancy within copies – the greater the chance a file will not disappear. But archives should not get involved in these complex interrelations.

Bandwidth has to do with the service requirements of the archive: how many users, how many concurrent users, where they are and how they are connected. It should be noted that for online archives with a large number of users (eg the general public), the bandwidth costs may far outweigh the storage costs—to the extent that the estimated storage cost could be less than the expected error in the estimated bandwidth costs, in which case storage is effectively (or comparatively) free.

5.3.   Storage media requirements

This paper will say nothing about storage media requirements. The point of view so far has been that what matters is the service. If a storage service provider can pull together a service based on magnetic tape, or minidiscs, or surplus 8" floppies from the 1970's, or holographic or molecular or optical tape storage or even by bouncing data to the moon and storing it in the delay time – it simply doesn't matter to the archive. All that matters is the storage is persistent, big enough, fast enough – and that when the files come back they can be used (currency). The fascination with storage media can be left to the storage industry. It's not the business of archivists.

5.4.   Repository Requirements

A digital archive will need to perform the following operations (at least):

- Acquisition:
  - For new material: bring files into the digital archive
  - Legacy material: digitisation from physical items to files
- Documentation:
  - An archive travels on its catalogue. As archives 'go digital', the catalogue becomes the major *value-added service* of the archive.
- Viewing:
  - The archive will have to support a multiplicity of 'proxies', because bandwidth will be insufficient to move high-resolution video files as quickly as MPEG-4 (or whatever) viewing files
  - Catalogue search, viewing and rough edit will, ideally, be combined in a single asset-management application
- Re-Use
  - Full-quality material will have to be delivered, as files, to edit suites or wherever else they are needed.
- Asset Management and Life-cycle management
  - There is a set of *birth to death* processes here, based on processes established in the document management world (where they started 'going digital' 20 years ago). Principal issues include access control, version control and digital rights management.

The functionality just listed is common to library / archive systems in general. They all have modules for acquisition, cataloguing and circulation control. There are two basic differences between a conventional library IT system and a digital repository:

- **a repository holds the content**, not just the catalogue and support for acquisition, circulation control and other processes
- a repository **prevents loss of content**, or at the least tries very hard to prevent loss – by incorporating processes and technology specifically aimed at insuring the continued viability (persistence and currency) of the content

At first glance, a conventional library system plus a bit of computer storage might look like a digital repository. The library system controls the functions (processes), and then instead of books on shelves it's just a matter of files in the storage. The whole difference between the "catalogue plus files" approach, and a reliable repository, is that the basic unit in the storage is much more that just a file (of text or digitised audiovisual signals or other content). The basic unit of a repository is an "information package". The package contains the content, but also contains **everything** else deemed necessary to understand the content, access the content (open or view or play the file) – and keep the content viable indefinitely.

Repositories need files, and information about files – that's how they differ from just files in mass storage – and that's how they can even have a hope of ensuring preservation of content. For this reason, the OAIS standard discussed in the next section doesn't talk about files – it has Submission Information Packages going in, and Dissemination Information Packages coming out. The central idea is that files of content need explanation and support – they need additional information – in order to survive and remain useful, and that a repository will acquire, store, and deliver *information packages* in order to keep the content viable.

6.   Repository Standards

A digital repository is defined in various ways: by its hardware and software, and by its operations (functionality). Because the functionality will need to survive the obsolescence and replacement of all its hardware and software constituents, the existing formal definitions/exemplars of digital repositories are defined principally in terms of what they do – their rules of operation.

The following section, Chapter 7, will discuss actual working repository software. This section describes standards for digital repositories – and there really only is one: the Open Archival Information System[46]: OAIS, which is also ISO standard 14721:2003[47].

The standard covers functionality and required information. Although it has a lot of detail, the document overall is at the conceptual level, rather than being a technical specification.

There could be many ways to implement OAIS, but the primary example of a true technical specification and implementation of the *information packages* of OAIS is the METS specification. OAIS and METS are discussed in the following two sections.

6.1.   OAIS

This is a project of the Research Libraries Group[48], an international, not-for-profit membership organization of over 150 universities, libraries, archives, historical societies and other institutions with collections for research and learning. Though based in North America, it has a global influence – in the library world.

OAIS, the Reference Model for an Open Archival Information System, is a comprehensive logical model describing all of the functions of a digital repository. It outlines how digital objects can be prepared, submitted to an archive, stored for long periods, maintained, and retrieved as needed—without addressing specific technologies or archiving techniques.

Thus OAIS is a model only, and says nothing about how to make it work. However such a model is very useful, because it will guide system development (if people follow it!), helping to ensure that systems have all the necessary functionality to really provide trusted, sustainable access.

The official description of the ISO Standard covering the OAIS model follows:

Abstract

ISO 14721:2003 specifies a reference model for an open archival information system (OAIS). The purpose of this ISO 14721:2003 is to establish a system for archiving information, both

digitalized and physical, with an organizational scheme composed of people who accept the responsibility to preserve information and make it available to a designated community.

This reference model addresses a full range of archival information preservation functions including ingest, archival storage, data management, access, and dissemination. It also addresses the migration of digital information to new media and forms, the data models used to represent the information, the role of software in information preservation, and the exchange of digital information among archives. It identifies both internal and external interfaces to the archive functions, and it identifies a number of high-level services at these interfaces. It provides various illustrative examples and some "best practice" recommendations. It defines a minimal set of responsibilities for an archive to be called an OAIS, and it also defines a maximal archive to provide a broad set of useful terms and concepts.

The OAIS model described in ISO 14721:2003 may be applicable to any archive. It is specifically applicable to organizations with the responsibility of making information available for the long term. This includes organizations with other responsibilities, such as processing and distribution in response to programmatic needs.

There are now various guides to OAIS. A paper by David Giaretta from the UK Digital Curation Centre[49] is particularly enlightening, at least to people who find diagrams in UML (Universal Modelling Language) enlightening. For instance, it opens an information package to show the contents and relationships:
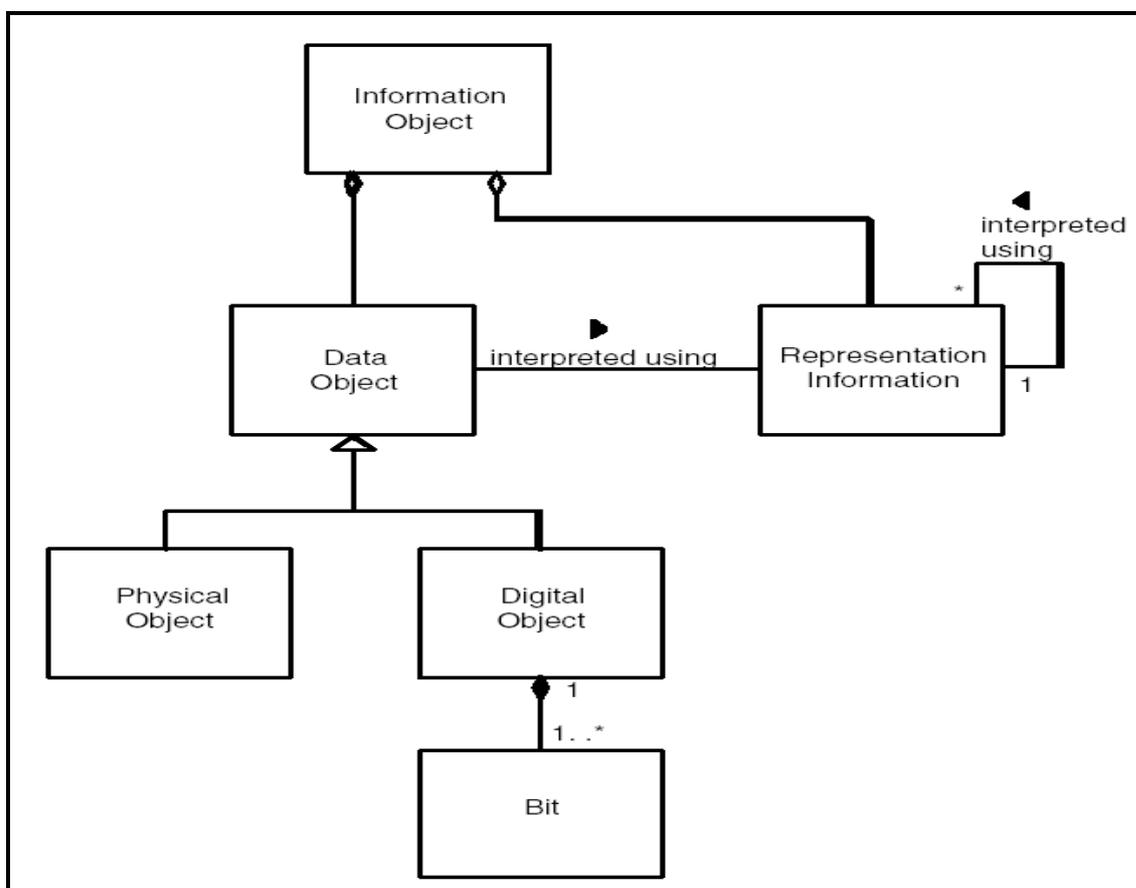
Figure 2 - OAIS Information Object

This UML diagram means that
- an Information Object is made up of a Data Object and Representation Information
- A Data Object can be either a Physical Object or a Digital Object . An example of the former is a piece of paper or a rock sample.
- A Digital Object is made up of one or more Bits .
- A Data Object is interpreted using Representation Information
- Representation Information is itself interpreted using further Representation Information

Institutions that are implementing OAIS include:

- **Cedars Project (CURL Exemplars in Digital Archiving; UK)**
  Cedars Guide to the Distributed Digital Archiving Prototype (pdf)

- **Harvard University's Digital Repository**
  Harvard University's Submission Information Package for electronic journals
  (June 2002)

- **Koninklijke Bibliotheek (Netherlands)**
   High-level design of a deposit system for electronic publications (NEDLIB Project)

- **Library of Congress**
  The Library of Congress's Archival Information Package for audiovisual materials
  (Presentation, June 2002. Multiple schematics within.)

- **Massachusetts Institute of Technology (MIT)** DSpace

- **OCLC Digital Archive** OAIS to OCLC Digital Archive

6.2.   METS

As mentioned, OAIS doesn't provide a technical specification, it provides a higher level framework. METS[50] gives the next layer, as described in the paper "Like Russian Dolls: Nesting Standards for Digital Preservation" by Günter Waibel of the Research Libraries Group.[51]

> "The **Metadata Encoding and Transmission Standard** (METS), developed by the library community, provides a data structure for exchanging, displaying, and archiving digital objects. It nests within the larger framework of the OAIS as a possible mechanism for data transfer between entities inside and outside the OAIS archive."

This whole paper has been arguing that, for long-term preservation, just putting files into storage is inadequate; more information, structure and control is needed. OAIS provides the structure and control, and defines – in outline – the information. METS is a detailed specification of the OAIS *information packages,* and so defines an exchange format for repository *information packages.*

Dr Waibel's paper continues:

> "Any community interested in implementing the OAIS has to identify or create a file-exchange format to function as an information package. For the cultural

heritage community, METS shows great potential for filling that slot. METS wraps digital surrogates with descriptive and administrative metadata into one XML document. Digital surrogates in this context could be digital image files as well as digital audio or video. At the heart of each METS object sits the structural map, which becomes a table of contents for public access. The hierarchy of the structural map allows the navigation of media files embedded in, or referenced by, the METS object. It enables browsing through the individual pages of an artist's book as well as jumping to specific segments in a time-based program, for example, a particular section of a video clip".

OAIS and METS are the main standards for digital repositories – but then there is implementation. The next section covers IT systems that are actually available and describe themselves as digital repositories – those which include OAIS and METS and other standardised approaches, and some which do not.

7. Repository Technology

Several computer systems arising from the university, national library and research library sector that attempt to fulfil the requirements of a digital repository. They have been developed as products, open source or commercial, and several have from dozens to hundreds of users.

These principal examples will be described in detail in the following sections
- the digital library model of the US digital libraries project: Fedora
- the HP-MIT open source product: DSpace
- Greenstone
- LOCKSS and EPRINTS

Further general information and review of digital repository technology is available from the UK academic library sector organisation UKOLN[52], and from the UK academic infrastructure provider JISC[53]. A website dedicated to tools and technologies for "open repositories" is i-tor[54] (in Dutch, but with open source development on SourceForge in English[55])

Major computer companies are also involved. HP is a co-developer of DSpace, and other work began with digital library systems in the early 1990's. IBM was an early developer, but has had no system specifically identified as a "digital library" product since 1997[56].

SUN also was an early developer of digital library technology, and had a digital repository project in Austria / Germany in 2004. The website appears dormant57.

Microsoft is a prime partner in the British Library's digital repository project.[58]

The Digital Preservation Coalition has a Directory of Digital Preservation Repositories and Services in the UK[59].

Finally, one very attractive proposition is to pay for a service rather than a product, and the international library organisation OCLC offers a repository service, described in Section 8,

The final section of this chapter looks at the relative costs of storing documents digitally rather than as paper on shelves, or as microfilm.

One survey of repository software[60] showed several more open-source projects or products. Such details as are available are given in the reference.
The systems are: Archimede, ARNO,CDSware, MyCoRe and OPUS.

7.1.   Fedora

Fedora is the Flexible Extensible Digital Object and Repository Architecture project of Cornell University (New York) and the University of Virginia, both in the USA[61].

Fedora is a general-purpose digital object repository system that can be used in whole or part to support a variety of use cases including: institutional repositories, digital libraries, content management, digital asset management, scholarly publishing, and digital preservation.

Fedora goes beyond OAIS, in that it is real: it has developed software which is available now, in its second version.

The Fedora repository system is open-source software licensed under the Mozilla Public License[62]. The interface to the system consists of three open APIs that are exposed as web services (in line with JISC and really all academic / research thinking about large-scale integrated systems: that they manifest as web services).

The Fedora services are:

- **Management API (API-M)** – defines an interface for administering the repository. It includes operations necessary for clients to create and maintain digital objects and their components. API-M is implemented as a SOAP-enabled web service.

- **Access API (API-A)** – defines an interface for accessing digital objects stored in the repository. It includes operations necessary for clients to perform disseminations on objects in the repository and to discover information about an object using object reflection. API-A is implemented as a SOAP-enabled web service.

- **Access-Lite API (API-A-Lite)** – defines a streamlined version of the Fedora Access Service that is implemented as an HTTP-enabled web service.

A full description (and link to download!) is available at the referenced webpage. Fedora had 85 delegates at their 2006 Users' Conference[63], representing 50 institutions. It is open source, but has a full installation and support service from the commercial organisation VTLS[64]. Fedora can ingest packages in METS, and in the MPEG-21 DIDL[65] metadata format. A recent review says "Fedora's claim to OAIS compliance mainly centres around its ingest and dissemination functions"[66].

7.2.   DSpace

DSpace[67] is a joint project of MIT Libraries and the Hewlett-Packard Company. It intends to provide stable long-term storage needed to house the digital products of MIT faculty and researchers. The software is also available to other institutions as open-source, though there is a commercial venture to provide supported implementations.

In general, it is a solution to archiving and maintaining access to the results of research projects and similar academic activity. As such, it is not as general as the intentions of OAIS/Fedora. However, it is a working package in real use.

Various university departments in the UK are considering or have started implementing DSpace, including the Cambridge. US implementers include Columbia University, Cornell University, Ohio State University, the University of Rochester, the University of Toronto, and the University of Washington at Seattle.

DSpace advertises itself as providing the following functions:

- Capture and describe digital works using a custom workflow process
- Distribute an institution's digital works over the web, so users can search and retrieve items in the collection
- Preserve digital works over the long term

DSpace uses METS technology, and its OAIS compliance has recently been described as "patchy but increasing"[68].

7.3.   Greenstone

Greenstone[69] is properly a digital library product, and has nearly a 10 year history of development at the University of Waikato in New Zealand, and at associated centres. It is unique among digital library projects in that it has always maintained a strong focus on the low-cost, low-technology end of information storage and distribution, as well as supporting large projects. An early result of the Greenstone work was technology to produce a "digital library on a CD" – which would in turn run on MAC or PC desktop computers through use of Internet technology. The CD content could also be loaded to a web server for more general distribution. All of this work was accomplished in the late 1990's[70].

In recent years the Greenstone team have also worked with repository standards. Greenstone can now import and export using the METS standard[71]. – as well as having direct import/export compatibility with Dspace.

Greenstone is open source and has full information on their SourceForge wiki[72].

Greenstone is used by about 40 academic and public-sector digital libraries, plus three UN agencies and for 35-40 humanitarian collections produced by the Human Info NGO in Belgium. There is extensive training. Its world-wide take-up is indicated by the fact that the user interface is available in nearly 40 languages[73].

7.4.   LOCKSS and EPRINTS :

There are two further international projects which are now mentioned in 'trusted repository' discussions, but which focus principally on academic articles and journals. These documents are vital to universities and their associated libraries, but may be less relevant to audiovisual collections.

Eprints[74] is focussed on open access to research information. Therefore is has core technology around getting information from one researcher to another, via 'self-archiving' and simple, common metadata. The main metadata approach is OAI – the

Open Access Initiative. The author has tried not to mention this, but now has to. OAI is useful and well worth mentioning in any general discussion of open access to information – and particularly to research. However OAI has a very unfortunate tendency to be confused with OAIS, which is just unfortunate.

Eprints needs to be mentioned because of its significance in the library world. Eprints itself does not have specific technology to ensure long-term preservation of documents – which is where LOCKSS has been introduced.

LOCKSS is "lots of copies keeps stuff safe", and is definitely aimed at making sure content does not disappear. LOCKSS works in the context of international research, where 'lots of copies' and needed. It is relatively trivial to make multiple copies of e-documents. LOCKSS is an organised way to ensure that there will be multiple copies in multiple places.

Eprints plus LOCKSS is digital repository approaches that basically ignores OAIS and METS and complex 'information packages' – and relies an standardised metadata plus … lots of copies. This approach could be of use for audio files, which are not huge (compared to video).   But the LOCKSS approach ignores the problem of format obsolescence. If documents are in a format that can no longer be opened and correctly interpreted, it doesn't have much to have 'lots of copies' of the document. So the LOCKSS approach cannot offer real security against obsolescence.

Eprints technology has been reviewed by the UK Digital Curation Centre.[75]

LOCKSS, Eprints and Dspace are all reviewed in an article[76] from Ariadne, the "Web magazine for information professionals in archives, libraries and museums in all sectors.

8.   Services, Costs and Conclusions

This section briefly considers the cost of digital repositories, looking at costs of a managed repository *service* rather than of a digital repository *product*. A product is essentially some hardware and some software, with a purchase price – but beyond that are the costs of the institution around the repository, with its staffing and maintenance that make the total system 'trusted', and give it any hope of permanence. Comparison is made with other forms of data management, particularly 'managed storage' – and with the raw costs of digital storage devices and the costs of shelf-based storage. The conclusion for print-based media is that repositories are hugely more expensive than the main alternative: microfilm. Audiovisual archives do not have the option of storage on microfilm. This section concludes with a recommendation for the **implementation of repository functionality (use of OAIS processes) in an off-line (tapes on shelves) rather than online (everything on servers) environment**.

Because digital repositories are both desirable and complicated, one major library organisation offers a 'trusted digital repository' service: the OCLC in the USA. The OCLC Digital Archive is a commercial proposition, and uses OAIS principles and the METS file exchange standard. As OCLC have been involved in developing these standards, it's not too surprising that they also implement them.

However the OCLC service is hardly going to become an overnight solution to storage and preservation problems for digital audiovisual collections, because of the relatively high costs. According to the 2002 figures quoted by Chapman[77], OCLC was charging US$15.00/GB per year (at the best rate, for a terabyte or more of storage). Further, that charge is basically just for secure digital storage, NOT for what Chapman refers to as "the capability to render intellectual content accurately, regardless of technology changes over time"[78].

The problem is: shelf storage had a cost in 2002 of about US$10 per shelf foot (for storage with full environmental control). At 10 videotapes per foot, one hour each at Digibeta quality (80 Mb/s), the storage is 400 GB for $4 (for the shelf) plus $40 for the tape itself, meaning about $0.10 per GB! The difference is a factor of 150.

Some further information on managed storage costs is given in the paper "The Digital Black Hole" by Jonas Palm of the Swedish National Archives[79]. Managed storage requires staff as well as equipment, and staff can be a major part of the total cost. Palm quotes a Microsoft source on information management costs (primarily from the text world, where 1 TB of data represents roughly 250 million pages). In the financial world 1 TB of data requires one system manager, and other costs add up to US$300k per terabyte. In other sectors this drops to one system manager per 10 TB, and for 'aggregators' like Google where data management is very automated, it drops again to one manager per 100 TB. Personal communication[80] with the Sun "Honeycomb" storage management project[81] also indicated a 'bottleneck at around 10 TB', with 'negative economies of scale' – meaning costs increased more rapidly as storage increased – for storage volumes greater than 10 TB.

These cost comparisons are summarised in the following table. The first row, 2002 costs, shows that managed storage is 150 times the price of storage of videotape on shelves, as discussed above (Chapman data). The middle column is for storage on hard drives of the cheapest sort – just the cost of the drives and nothing more.

Year	Analogue on shelves	Digital media (offline)	Managed storage (online)
2002, cost for 1 GB	$0.10	$4	$15 = 7 + 8
2006 (estimate)	$0.11	$1	$11 = 7 + 4
2010 (estimate)	$0.12	$0.25	$9 = 7 + 2
2020 (estimate)	$0.15	$0.02	$7.05 = 7 + 0.06

Table 1- Estimates of audiovisual storage costs

As shown in the successive rows of the table, cost in the middle column drop dramatically. This drop is the storage variant of Moore's Law[82]: storage capacity (for the same cost) doubling every 24 months. This drop has been steady for about 40 years. Prospects for the next 10 years are covered in another PrestoSpace publication: "Ten-year Forecast of Storage Evolution"[83]

The managed storage figures quoted by Palm indicate that around US$30k per terabyte, meaning US$30 per gigabyte, was the "going rate" in 2005. These figures are higher that the Chapman figures, because they refer to IT industry in general rather than OCLC specific service for archive storage. But the difference between the Chapman and the Palm figures does not matter. What matters is:
- the figures are in rough agreement;
- the real issue is what happens to storage costs in the future.

Table 1 starts with the Chapman figure of $15/GB/yr, and looks at what happens as Moore's Law applies to the raw storage part of that figure. We assume that about half the cost in the Chapman / Palm data represents staff and facilities[84], not storage devices themselves. These costs do not drop, but neither do they rise. Small improvements in IT management technology offset inflation, and basically IT systems cost now what they did 20 years ago – but the amount of storage in the systems (for that same price) has grown by the storage variant of Moore's Law[85]: storage capacity (for the same cost) doubling every 24 months.

As seen in the right-hand side of the table, raw storage costs drop by Moore's Law, and all that is left, eventually, is the rest of the cost: management, maintenance, facilities. The table shows managed storage costs levelling off, not dropping as in the middle column. The numbers in the right-hand column could be off by a factor of two, but the general conclusion is that managed storage will NOT continue to reduce in cost. There will be little reduction after 2010, with cost stabilising at about $10/GB/yr.

The first column, shelf costs, increases only as general inflation increases – which in most areas that are using digital repositories is quite low. The middle column follows Moore's law and shows storage becoming practically free – but only for raw storage devices, not managed servers.

The implications are very plain:
- archives will continue to find a managed service unaffordable, if by managed service one means a fully online system of mass storage (server-based storage; mass storage);
- letting materials sit as they are on shelves is untenable, as digitisation is required for preservation – certainly for audio and video materials;
- what's left is the middle column: use of raw storage media, which offers digital storage and low cost.

The problem with raw storage media is that it simply can't be trusted – for three years much less the indefinite future. Raw media is has NONE of the attributes of a trusted repository, and would fail all the items in the US and UK checklists mentioned in Section 2.2. Material just dumped on storage media is very much at risk. The media is prone to failure, there is no inherent way to find anything, and the files themselves if they can be found and read (as data) will become obsolete and unplayable – and possibly unidentifiable and, effectively, lost.

The need, and really the only hope for the huge amounts of digital storage required by audiovisual archive, is to combine the physical storage approach of the middle column – offline storage – with the functionality associated with the highest levels of service from online trusted digital repositories. Clearly what is needed are **trusted offline repositories,** with the security of the online repositories of the right-hand column of Table 1, at costs closer to the centre column.

PrestoSpace will support archives in convincing the storage industry that what archives need is to *preserve content,* not simply to write files to storage.

The PrestoSpace conclusion is that archives need to use – and industry needs to supply – **OAIS-compliant methodology for offline storage**. There needs to be a mechanism for *submitting an information package*, not just writing a file. The offline storage must manage the whole information package. This is basically a library function: a formal acquisition procedure including the *life-support system* metadata which is the heart of the OAIS approach. The library system (asset management system) needs to be sufficiently robust to maintain the link between the file and the rest of the *information package*, indefinitely.

Archives and industry will need to focus on this issue, and there should be a Centre for defining the approach in detail, and supporting both industry and archives in the implementation. It is already difficult to introduce concepts like OAIS to the IT industry. For audiovisual material, these concepts have to been introduced to the post-production and facility house industry – because that is where the files and the *information packages* will be produced. It will take a lot of support, from a Competence Centre because there is nothing else suitable, in order to lead facility houses through the OAIS maze and into the realm of trusted repositories.

The need for *trusted offline storage* applies across digital libraries, but is especially urgent for audiovisual archives because of the size of audiovisual data. Other media can use online storage for their digitisation projects – though they then risk dropping into the black hole described in the Palm paper. Audiovisual archives cannot but full-quality video on mass storage in any significant volumes, and so need to move

directly to secure offline storage. *Trusted offline storage* will then provide a lifeboat for rescuing other digitised media from the black hole effect – that online storage is unaffordable. Offline storage will hold large files of high data rate media, and online storage can be reserved for user access to low data rate proxies. This model has been in use for a long time, but without the essential factor of making the offline storage **trusted** – using the OAIS approach to store full *information packages* instead of just files.

An essential aspect of an information package is identification of the content of a file, particularly with regard to understanding the file format. There is then the large issue of ensuring that formats remain viable. The important point, and a major motivation for a Competence Centre, is that format viability isn't ensured locally. Identification is a local requirement, but once the file type has been labelled, the rest of the processes needed to ensure format viability can be implemented once, in one place. These processes include maintaining players for the format, and establishing a migration path to new formats and new players. This aspect of digital preservation is not an item by item issue, but a general industry issue that can be managed in one centre of sufficient competence – a Competence Centre.

The conclusions of this section are:
- offline digital media has a huge role in the future of audiovisual archives
- management of such media needs to move from current practice to the processes established for digital repositories.
- There is a strong requirement for a Centre of Competence to guide archives and industry, and provide – centrally – the technology for a migration path for audiovisual offline content. This is the key technology needed for what Chapman referred to as "the capability to render intellectual content accurately, regardless of technology changes over time" [86].

9. References:

[1] Competence Centres for digitisation and preservation exist in The Netherlands, and PrestoSpace has proposed an Audiovisual Competence Centre for supporting audiovisual digitisation and preservation across Europe.

[2] credited to: www.rlg.org quoted: http://www.bl.uk/about/strategic/glossary.html

[3] credited to: http://www.openarchives.org/OAI/openarchivesprotocol.html#Introduction
    quoted by: www.edtechpost.ca/pmwiki/pmwiki.php/Main/GlossaryAnalysis

[4] Anglo-American Cataloguing Rules, en.wikipedia.org/wiki/**AACR2** **and** www.**aacr2**.org

[5] www.rlg.org/longterm/repositories.pdf

[6] www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf

[7] www.ariadne.ac.uk/issue48/dempsey/

[8] www.searchtools.com/info/faceted-metadata.html

[9] www.minnesotatechnology.org/publications/Magazine/2004/Spring/HideandSeek.asp

[10] www.microsoftfrontpage.com/content/ARTICLES/GettingFound.html

[11] en.wikipedia.org/wiki/Web_portal

[12] en.wikipedia.org/wiki/Union_catalog

[13] mic.imtc.gatech.edu/unicatlg_1.htm MIC: A union catalogue for Moving Images

[14] www.biblio-tech.com/html/z39_50.html Z39.50, the standard underlying unification of library data

[15] en.wikipedia.org/wiki/Private_network Private Networks

[16] en.wikipedia.org/wiki/VPN Virtual Private Networks

[17] www.prestospace.org/project/public.en.html D15.3

[18] www.prestospace.org/project/public.en.html D15.1

[19] www.dpconline.org Digital Preservation Coalition

[20] www.**loc**.gov/rr/mopic/avprot/ Library of Congress Audiovisual Preservation

[21] www.digitalpreservation.gov/ US National Digital Information Infrastructure and Preservation Program

[22] www.nla.gov.au/padi/ subject gateway to international digital preservation resources

[23] www.erpanet.org/ Electronic Resource Preservation and Access Network

[24] www.digitalpreservationeurope.eu/ Digital Preservation Europe

[25] http://www.mtsu.edu/~kmiddlet/stateportals.html lists several hundred digitisation projects in the USA, and http://www.tasi.ac.uk/resources/casestudies.html gives 15 image digitisation case studies in the UK

[26] www.srlf.ucla.edu/exhibit/default.html "The **History of Microfilm**: 1839 to the Present"
http://www.heritagemicrofilm.com/History.aspx
see also:
  - Dalton, Steve. Microfilm and Microfiche. Technical Leaflet Series, 5, No. 1. Andover, MA: Northeast Document Conservation Center, 1999.
  - Online Tutorial Local Government Records: Just the Basics . Lesson 8: Microfilm. Columbus, OH: Ohio Historical Society, 1998.
  - Preservation Microfilm Bibliography. Bethlehem, PA: Preservation Resources/OCLC, 1998.
  - RLG Archives Microfilming Manual. Mountain View, CA: Research Library Group, 1994. 208p.

[27] www.**corbis**.com/

[28] www.creative.**gettyimages**.com

[29] www.apple.com/**itunes**/

[30] www.**youtube**.com/

[31] www.library.cmu.edu/Libraries/MBP_FAQ.html and www.archive.org/details/**millionbooks** Million Books Project

[32] www.**opencontentalliance**.org

[33] www.archive.org/

[34] three.co.uk/news/h3gnews/pressnewsview.omp?collcid=1019745742912&cid=1163002510792&index=7

[35] en.wikipedia.org/wiki/**High-definition_television**

[36] www.smh.com.au/articles/2005/09/21/1126982069009.html

[37] www.brainyquote.com/quotes/quotes/m/marktwain141773.html

[38] SAM: prestospace-sam.ssl.co.uk ; PrestoSpace: www.prestospace.org

[39] http://www.nla.gov.au/padi/topics/36.html

[40] Frank McCown, Sheffan Chan, Michael L. Nelson, and Johan Bollen (2005). "The Availability and Persistence of Web References in D-Lib Magazine". *Proceedings of the 5th International Web Archiving Workshop and Digital Preservation (IWAW'05).*

[41] http://www.w3.org/Consortium/Persistence

[42] PRONOM: nationalarchives.gov.uk/pronom/ ; LOC Digital Formats: digitalpreservation.gov/formats/

[43] UKOLN: www.ukoln.ac.uk/interop-focus/gpg/Preservation/ ; PADI: www.nla.gov.au/padi

[44] www.nla.gov.au/padi/

[45] PrestoSpace: www.prestospace.org
Sam: http://prestospace-sam.ssl.co.uk/
Ready-Reckoner: http://prestospace-sam.ssl.co.uk/

[46] www.rlg.org/en/page.php?Page_ID=3201 OAIS

[47] nost.gsfc.nasa.gov/isoas/ref_model.html

[48] www.rlg.org Research Libraries Group

[49] dev.dcc.ac.uk/twiki/bin/view/Main/DCCApproachToCuration

[50] www.loc.gov/standards/mets/

[51] http://216.239.59.104/search?q=cache:8sIsWsLR6H8J:www.rlg.org/preserv/diginews/v7_n3_feature2.html+OAIS+mets&hl=en&gl=uk&ct=clnk&cd=7

[52] www.ukoln.ac.uk/metadata/resources/digital-repositories/

[53] www.jisc.ac.uk/uploaded_documents/digital-repositories-review-2005.pdf

[54] http://www.i-tor.org/nl/

[55] http://sourceforge.net/projects/i-tor

[56] http://www-306.ibm.com/common/ssi/OIX.wss?DocURL=http://d03xhttpcl001g.boulder.ibm.com/common/ssi/rep_ca/1/897/ENUS204-211/../../../0/897/ENUS204-180/../../../8/897/ENUS204-178/../../../9/897/ENUS204-179/../../../7/897/ENUS104-117/../../../1/897/ENUS102-271/../../../1/897/ENUS202-031/../../../8/897/ENUS201-248/../../../2/897/ENUS200-352/../../../5/897/ENUS200-355/../../../0/897/ENUS200-030/../../../5/897/ENUS297-355/../../../2/897/ENUS297-312/index.html&InfoType=AN&InfoSubType=null&InfoDesc=Announcement%20Letters&paneltext=&panelurl=&singlehitflag=false&printableversion=yes

[57] www.coe.hu-berlin.de/

[58] www.dpconline.org/docs/events/0610drambains.pdf

[59] www.dpconline.org/docs/guides/directory.pdf

[60] www.soros.org/openaccess/pdf/OSI_Guide_to_IR_**Software**_v3.pdf

[61] www.fedora.info/ Fedora

[62] www.mozilla.org/MPL/MPL-1.1.html Mozilla

[63] www.lib.virginia.edu/digital/fedoraconf/

[64] www.vtls.com/

[65] xml.coverpages.org/**MPEG21**-WG-11-N3971-200103.pdf Digital Object Description Language

[66] www.ukoln.ac.uk/projects/grand-challenge/papers/**oais**Briefing.pdf

[67] libraries.mit.edu/dspace-mit/ and www.dspace.org/ DSpace

[68] www.ukoln.ac.uk/projects/grand-challenge/papers/**oais**Briefing.pdf

[69] www.**greenstone**.org

[70] portal.acm.org/citation.cfm?id=336650&coll=portal&dl=ACM "Greenstone: a comprehensive open-source digital library software system" Proceedings of the fifth ACM conference on Digital libraries San Antonio, Texas, United States ; Pages: 113 - 121 (2000) ISBN:1-58113-231-X

[71] www.greenstone.org/cgi-bin/library?e=p-en-home-utfZz-8&a=p&p=factsheet

[72] www.**greenstone**.sourceforge.net/**wiki**/index.php/**GreenstoneWiki**

[73] www.greenstone.org/cgi-bin/library?e=p-en-home-utfZz-8&a=p&p=factsheet

[74] www.eprints.org/

[75] www.dcc.ac.uk/resource/technology-watch/eprints/

[76] www.ariadne.ac.uk/issue43/prudlo/

[77] jodi.tamu.edu/Articles/v04/i02/Chapman/

[78] jodi.tamu.edu/Articles/v04/i02/Chapman/ p4

[79] www.tape-online.net/docs/**Palm_Black_Hole**.pdf

[80] Personal conversation with Mike Davis, **Honeycomb** senior project manager. Sun Honeycomb have participated in PrestoSpace STAG (Storage Technology Advisory Group) events

[81] research.sun.com/minds/2005-0628/

[82] en.wikipedia.org/wiki/Moore's_law

[83] www.prestospace.org/project/deliverables/D12-5.pdf

[84] This cost breakdown is deduced from the "one staff member per ten terabytes" statement quoted by Chapman. If storage costs around $100k per $300k for 10 TB, and includes one full-time staff member, then it is reasonable to conclude that about half the cost is for the member of staff and for all associated infrastructure.

[85] en.wikipedia.org/wiki/Moore's_law

[86] jodi.tamu.edu/Articles/v04/i02/Chapman/ p4