Module 3.1: Metadata to manage digital information

Introduction

Although you might not realise it, we are all users of metadata all the time. The data about individual objects in the collection that is the basis of library catalogues is one example with which you are probably familiar. But metadata is all around us, and although often overlooked, is a common feature of everyday life – think about the labels on bottles and cans of drink for instance, they are covered in metadata! Metadata is all around and an inescapable part of everyday life. You need to look no further than the recent controversy in Australia about the Federal Government's proposal (now legislation) to retain all the metadata created and collected by ISPs for two years, to see not only how much the topic is misunderstood but also how important it is.

Think, too, about the discussion that ensued when it became clear that the Government and its bureaucrats had no idea what metadata is or how powerful it is, and the many ways that metadata on its own (i.e. without the content to which it refers) can be analysed and used. Metadata, commonly but incompletely defined as ‘data about data’ is an essential part of managing digital objects of any type, but especially digital records. Just as metadata is crucial to understanding and using any records, it is indispensable for ensuring the authenticity of digital records as well as keeping them accessible and useable over time. This topic explains metadata and examines some of the uses to which it is put in digital preservation.

Types of metadata

A commonly seen definition of metadata is that it is ‘data about data’. While this is short and snappy, and strictly correct, it is not very helpful and rather uninformative. There are many definitions of metadata out on the web if you care to look, and they all have a lot of similarity, but one that we will use for the purposes of this module is found in the National Archives of Australia’s online glossary:

Structured information that describes and/or allows users to find, manage, control, understand or preserve other information over time …

You may have heard of some metadata schemas: for instance, Dublin Core (DC) or AGLS (DC-based). These are examples of resource discovery metadata, whose aim is to assist in identifying information resources and guide people to them. And there are many other metadata schemas used by all sorts of communities of practice in all sorts of disciplines and areas of human activity. A common but not very helpful categorisation of metadata that you will come across (if you haven’t already) describes three categories of metadata:

- descriptive: for resource discovery and access
- structural: how compound resources are structured
- administrative: for managing resources over time.

This view of metadata types is somewhat inadequate for many reasons but principally because all metadata is ‘descriptive’ (i.e. it is always used to describe the characteristics of something) and making a separate category for descriptive metadata loses the point. As well, structural is a very specific and narrow category, while administrative is very broad and non-specific. Such a categorization does not help us understand metadata’s place and its uses – in which of the above categories do you put ‘preservation metadata’, for example. It might be better to categorise metadata schema by the use cases they serve, such as:

- Data management (e.g. ISO 11179)
- Privacy (e.g. W3C Platform for Privacy Preferences)
- Intellectual Property Rights (e.g. INDECS)
- Geospatial (e.g. ISO 19115)
- Records management/recordkeeping (e.g. ISO 23081)
- Resource Discovery (e.g. Dublin Core (DC), AGLS)
- Preservation (e.g. PREMIS)

It is important to realise that these different uses and the schemas that support them are not mutually exclusive – they are different and aimed at different purposes but they are complementary, i.e. it is not necessarily a case of choosing one over another. So while you might prefer AGLS over DC, using either does not then preclude the use of ISO 11179 or ISO23081, etc. Note that the categorisation discussed above is not really an important issue and we will not go into it further here.

However, while the sort of categorisation discussed above is probably useful mainly for identifying schemas available to meet your particular use case, it is important to know that there is another major distinction to be made between **metadata content standards and metadata encoding standards**. Content standards (also known as semantic standards, i.e. they are about meaning) set out the specific characteristics of an information object which are being described (such as the **author** property in DC and AGLS, for example), while encoding standards (also known as syntax standards) specify the actual machine-readable encoding for expressing semantic metadata and are aimed at machine readability and online network interoperability. Examples of each include:

- **Content standards**: AACR2, ISAD(G), ISAAR(CPF), DACS, RAD, MAD, DCMES, AGLS, ISO 23081, NAA Recordkeeping Metadata Standard
- **Encoding/syntax standards**: MARC, EAD, EAC, XML, HTML.

Thus you can use ISAD(G) as a recordkeeping metadata content (semantic) standard and use EAD as your syntax standard to express or write out the ISAD(G) metadata in order for it to become machine-processable.

**Preservation metadata**

You should be able to recognise that preservation metadata is a specific type of metadata serving a specific use case. Think about what the aim of preservation metadata might be – what is the use case?

In some sense, **preservation metadata is a sub-set of recordkeeping metadata**, because it is a necessary part of certain activities involved in managing digital records over time. These activities are the **recordkeeping processes needed to ensure digital records remain accessible and useable over time and across changing technical environments** – they include knowing the technical environment in which the record was created, and documenting any and all preservation actions (such as migration) that are applied to the record, as well as any authorised changes that have been made to the record.

**Preservation metadata has multiple functions**, including identifying the digital resource that is being preserved, recording what happens to the digital resource over time (its history), recording its context (who created it, why, how it was used, and so on), and providing sufficient information about the object so that future users and preservers of the resource have the information they need to ensure that the resource can be re-presented to the user in the future, despite changes in technology. All of this preservation metadata needs to be maintained over time - often, for a very long time. It also needs to be securely linked with the digital resource in such a way that the links between them can't be broken.

If this sounds like a tall order, you are right. Preservation metadata schemas are, of necessity, complex. They consist of many more elements than a typical resource discovery metadata
schema and approach existing recordkeeping metadata schemas (such as the NAA Recordkeeping metadata standard) in depth and complexity.

This is not the place for a history of preservation metadata development but it is sufficient to say that a number of early digital preservation initiatives realised the necessity of creating preservation metadata and started the development of preservation metadata schemas. Early work on such schemas was carried out by the National Library of Australia, the European NEDLIB project and the CEDARS project at the Universities of Cambridge, Leeds and Oxford in the UK.

Activity
Browse this online publication from the Getty Institute in the US: Introduction to Metadata, version 3.

Lavoie, Brian & Richard Gartner (2013). Preservation metadata (2nd ed.). Look through this document - it is comprehensive and wide ranging so do read the introduction, and at least browse the rest.

Preservation metadata standard

PREMIS

In March 2000 in the US, OCLC (a not for profit library services institution) and the Research Libraries Group (RLG) started a cooperative effort to encourage development of the necessary infrastructure to secure long-term retention of digital materials. A major focus of their work was to promote consensus in best practices for the use of metadata to support digital preservation. In 2001-2002, OCLC and RLG convened an international working group of experts to develop a metadata framework to support long-term retention of digital materials. In June 2002, this working group produced an important report on preservation metadata, A Metadata Framework to Support the Preservation of Digital Objects, which included a comprehensive survey of the types of information included in preservation metadata.

In June 2003, a second working group was formed to address implementation issues associated with preservation metadata and to further progress the recommendations of the 2002 report. This was the working group called PREMIS (Preservation Metadata Implementation Strategies) and which published, in May 2005, the first version of the PREMIS Data Dictionary under the title Data Dictionary for Preservation Metadata: Final report of the PREMIS Working Group. The PREMIS Editorial Committee, which maintains the PREMIS metadata standard, is now managed by the US Library of Congress (see below). The Data Dictionary is now up to version 3 (2015), with version 2.2 having been published as the second edition in 2012. For a good overview of the function of a data dictionary, check out this entry in Wikipedia, Data Dictionary.

The original PREMIS working group was an international working group with representatives from all over the world (including Australia and New Zealand) and the Data Dictionary is therefore based on widespread international consensus. It has become widely accepted as the definitive standard for preservation metadata.

The PREMIS Data Dictionary

The Data Dictionary (the PREMIS schema in practice) is a practical resource for implementing preservation metadata in a digital archive. It provides a “Canonical” definition of preservation metadata as “the information a repository uses to support the digital preservation process”. The
Data Dictionary is implementation independent, i.e., does not define how the metadata should be encoded or stored.

The simple data model (illustrated to the right) developed by PREMIS defines five types of entities which were seen to be particularly important for digital preservation:

- **Intellectual entity**: a set of content that is considered a single intellectual unit for purposes of management and description: for example, a particular book, map, photograph, or database. An Intellectual Entity can include other Intellectual Entities; for example, a Web site can include a Web page; a Web page can include an image. An Intellectual Entity may have one or more digital representations.
- **Object**: a discrete unit of information in digital form
- **Environment**: Technology (software or hardware) supporting a Digital Object in some way (e.g. rendering or execution). [Note that Environment is more of a sub-entity within Object]
- **Event**: an action that involves or impacts at least one Object or Agent associated with or known by the preservation repository
- **Agent**: person, organization, or software program/system associated with Events in the life of an Object, or with Rights attached to an Object
- **Right**: assertions of one or more rights or permissions pertaining to an Object and/or Agent.

[Taken from the PREMIS Data Dictionary, version 3.0]

It is also worth noting that PREMIS deals with three types of Objects:

- **File**: an actual file on an operating system (e.g. a PDF file)
- **Bitstream**: a series of bytes (1s and 0s) within a file which have meaningful properties in themselves (e.g. the header information within a JPEG2000 image file)
- **Representation**: a set of files (including structural metadata) which are required to render a single intellectual entity (e.g. a webpage consisting of HTML, CSS, and images, all necessary to render something useable)

The Data Dictionary then sets out what it calls **semantic units** for the four primary entities it describes. Semantic units are so called because this is a data dictionary, not a schema, but they correspond to what you probably know as metadata elements or properties. Semantic units are ‘properties’ of the entity to which they are attached, so ‘size’ is a semantic unit for object, providing information about the size of an object in bytes, e.g. ‘450678’ (NB. Because the semantic
unit is defined as size in bytes no unit of measurement is given).

There is a large number of semantic units most of which also have semantic components, which are more granular semantic units (properties) that are used to make up the information needed for a semantic unit. For example, size is a semantic component of the semantic unit objectCharacteristics, for the object entity. The objectCharacteristics semantic unit also contains a number of other components, e.g. fixity, format, and creating application (among others). In some cases, semantic components themselves have semantic components, so fixity, for example, has the semantic components messageDigestAlgorithm, messageDigest, and messageDigestOriginator. These nested semantic units can be as deep as 4 levels, so you can see that PREMIS does have a certain amount of complexity that can make it less than easy to implement.

Nevertheless, PREMIS can be used in a fairly straightforward manner and the depth of complexity does not have to be implemented in its entirety. Like other metadata standards, PREMIS has a certain core set of mandatory semantic units which must be used for compliance but many of the semantic units are optional, or only used under certain conditions.

Activity
You will find it interesting to browse the PREMIS home page. Some of the material is quite technical but there is also general material you should find accessible and interesting, as well as a wiki for PREMIS implementers which is worth looking at. Make sure you look at the Data Dictionary; even if you don’t read all of it at least understand how it is structured and what the five entities are.

PREMIS Issues
Although the PREMIS schema represents world’s best practice in preservation metadata, and while it is the predominant preservation metadata schema in use globally, it is not without problems. We don’t have time or space in this module to explore these issues in depth but it is worthwhile taking time to think about the issues raised here and to try to understand their consequences. One general issue that is still being grappled with is how do we know what preservation metadata people require to access, understand and re-use a digital object in the future? Digital preservation initiatives are still attempting to understand what is the minimum amount of preservation metadata necessary to carry a useable digital object forward through time.

More specific issues with PREMIS include:

- Number of semantic units is very large which can make it difficult to implement
- Does not include metadata properties to assist discovery
- Does not define the specific characteristics of Agents
- Does not directly consider rights and permissions not directly associated with preservation actions, e.g. access or reuse rights associated with the object itself
- Does not deal with technical metadata for all different types of digital file (left to format experts)
- Does not deal with the detailed documentation of media or hardware (left to media and hardware specialists)
- Does not consider resource management.

Although the Data Dictionary says the properties defined in PREMIS are the “things that most working repositories are likely to need to know in order to support digital preservation”, in fact this is only a restricted set of the metadata needed to ensure authentic digital records are carried across time and technological change. A lot of the properties archivists think necessary for recordkeeping are absent from PREMIS and it is very important to understand that while using
PREMIS metadata will help ensure digital records are carried forward through time, it is not on its own sufficient to guarantee the authenticity, reliability, useability and integrity of digital records over time.

METS

METS (Metadata Encoding and Transmission Standard) deserves a mention in this section because, although it is not a preservation metadata standard, it plays an important role in ‘packaging’ preservation and other metadata with a digital resource and allowing the resource to be shared by more than one repository. Michael Day describes METS as ‘essentially... an XML-based framework in which different types of metadata can be packaged together’: this is of value for digital preservation because it ‘can support the management and exchange of digital library objects' (Day, 2005, p.25). METS is maintained by the Library of Congress, and is being developed as an initiative of the Digital Library Federation. (XML is discussed in Module 2.3.)

A METS document has seven sections:

- a METS Header for brief descriptive information about the METS document itself;
- Descriptive Metadata;
- Administrative Metadata (which includes technical information about the file, information about intellectual property rights, the source material, and provenance metadata that records relationships between files and migrations);
- a File section listing all of the files that make up the object;
- Structural Map and Structural Links sections that enable individual files and metadata to be mapped to the structure of the object;
- a Behavior section that provides information on how particular components should be rendered.

METS is a useful tool as a container for all sorts of metadata and digital objects. In one XML package the metadata (structural, administrative and rights metadata as well as preservation metadata) can be combined with the digital object. METS is open standard and non-proprietary, and has been developed by the digital library community for that community. It has 'extension' metadata schema, one of which is PREMIS for preservation metadata. Because METS is flexible and expandable it is possible to incorporate metadata from other metadata schema (such as recordkeeping metadata) within a METS package. This makes it an extremely promising and useful packaging standard for digital preservation purposes.

Activity
Browse The METS web page containing documentation, presentations about METS, software to help users implement and use METS, examples, a tutorial, and more. (You will need some knowledge of XML to understand the examples and make full sense of the tutorial.)

This is an interesting article about re-thinking the use of metadata in a specific repository context: The Functions of (Meta)Data: Lessons Learned with a Fedora Digital Repository.

Preservation metadata issues

One standard or many?

As noted above, PREMIS is based on widespread international consensus and has become widely accepted as the definitive standard for preservation metadata. However, although PREMIS has become almost the de facto international standard for preservation metadata it has not swept all before it and other work on preservation metadata work schema has carried on.
Other earlier attempts at defining preservation metadata schemas came out of the CEDARS project at the University of Leeds in the UK in 2000, and the NedLib Project at the French National Library, also in 2000. One of the most significant activities developing an alternative preservation metadata schema is work being undertaken at the National Library of New Zealand.

**NLNZ Metadata Standards Framework – Preservation Metadata**

The National Library of New Zealand was an early developer of a preservation metadata scheme, and has also developed a software tool for the automatic generation of metadata (noted earlier in this topic). Its preservation metadata schema is based on the work of the OCLC/RLG Working Group (later developed into PREMIS). It is described on its website in these words:

This preservation metadata schema details the data elements needed to support the preservation of digital objects and will form the basis for the design of a database repository and input systems for collecting and storing preservation metadata. It incorporates a number of data elements needed to manage the metadata in addition to metadata relating to the digital object itself. The aim has been to produce a document that will serve as an implementation template while at the same time remaining consistent with standards being developed internationally around preservation metadata.

This schema consists of 18 elements describing the digital object, 13 elements to record the history of actions performed on the object, nine for technical information about file types, and five to record the history of changes made to the preservation metadata. It was first released in November 2002, and revised in 2003.

**Metadata creation**

The other major issue that is a problem for all metadata, not just preservation metadata, is the issue of metadata creation. Manual metadata creation takes time and, because it involves human resources, is notoriously costly compared to using technology (it is a ‘hand-crafted’ activity, to use a term common in discussion about digital preservation). This means that it would be highly desirable if preservation metadata could be automatically generated without human input, or more likely with only minimum human intervention.

However, tools for the automation of metadata creation are still relatively unsophisticated, and the field is continuously developing. Nevertheless, automated metadata creation is really the only viable approach given the huge quantities of digital objects we will be managing in the future, and is a key to effective digital preservation.

**Metadata extraction tools**

Automation of metadata creation involves the extraction of the metadata that already exists within or linked to digital resources. There are now quite a few tools that will do this. One that has been around for a while is the National Library of New Zealand's metadata extraction tool, so named because it automatically extracts metadata from digital files. It examines files in a range of common file formats (MS Word (various versions), WordPerfect, Open Office, MS Works, MS Excel, MS Powerpoint, TIFF, JPEG, WAV, MP3, HTML, PDF, GIF, and BMP) and extracts preservation metadata from these files. This metadata is output in XML format. This metadata extraction tool, as well as information about it, is available from the National Library of New Zealand's website.

Another tool that has been around for a while is the JHOVE tool developed by Harvard University as a file format identification and validation tool. Along with identifying and validating the format of a digital object, JHOVE also extracts technical metadata from the object and outputs it as plain text which can be readily processed into PREMIS metadata.
A third tool is the UK National Archives’ (TNA) DROID. DROID works with TNA’s PRONOM file format registry (a centralised format registry updated and maintained by TNA) and is primarily a file format identification tool. DROID is a local client which interfaces with the PRONOM database and in the course of identifying a digital object extracts relevant technical metadata from the PRONOM entry for the format. This metadata is output by DROID in an XML file.

The area of automatic metadata extraction certainly is an area that was neglected until recently. Nowadays, as you can see from the COPTR page below, there are a lot of tools around that will do metadata extraction. As the DCC noted sometime ago ‘as the amount of digital information being produced everyday increases at an exponential rate, it [manual creation of metadata] will eventually become impossible. This suggests that fully or partially automating this process will become essential’ (Ross & Kim). Researchers are attempting to use mathematical and computing techniques to examine questions about how to automate the process, and what the reliability of the result is. The field of metadata extraction/creation tools is rapidly developing and will bear close attention, if this is a topic that interests you.

**Activity**
Browse these resources:

- This comprehensive chapter on automated metadata generation from the Digital Curation Centre’s *Curation Reference Manual: Automated Metadata Generation*
- The COPTR webpage that lists metadata extraction tools: **Category: Metadata Extraction**.
- The [JHOVE web page](http://jove.sourceforge.net/) explains the use cases for JHOVE and provides links to other tools and guidance. JHOVE itself is free and available for download from Source Forge.
- DROID and PRONOM are described in brief on the UK National Archives website. More detail about DROID is on the DROID download page and the tool is available free of charge. A [video](http://jove.sourceforge.net/) from a 2012 Conference in New Zealand gives extra information about the development of DROID and PRONOM.
- In 2012, the National Library of New Zealand did some exploratory research on DROID & PRONOM and there is a video of some of their findings presented by Jay Gattuso which you might find interesting as it deals with the issue of persistence of file format identifications. The video is 23 minutes long.

**Checksums: a specialised kind of metadata**

Checksums are a specialised example of metadata, and one that you will come across frequently in literature and discussions of digital preservation. A checksum is a value computed from a digital file using a mathematical algorithm. This value is used to check for errors in the transmission or storage of the digital file by re-calculating the value and seeing if it has changed. If it is different some change has occurred to the bits in the original file.

Checksums can be explained by the analogy of ISBNs (International Standard Book Numbers). This is a 10 or 13 digit number that uniquely identifies the publication to which it is assigned. The last digit in the sequence is a check digit, calculated according to a formula which multiplies each of the first nine digits with a specified number, then divides the total by 11. The difference is the check digit. Here’s an example:

**check digit for an ISBN-10 of 0-306-40615-__ is calculated as follows**

\[
0 \times 10 + 3 \times 9 + 0 \times 8 + 6 \times 7 + 4 \times 6 + 0 \times 5 + 6 \times 4 + 1 \times 3 + 5 \times 2
\]

\[
= 0 + 27 + 0 + 42 + 24 + 0 + 24 + 3 + 10
\]

= 130

The check digit is 0, so the calculated checksum matches the actual checksum on the cover of the book. If it didn’t match, there would be an error in the transmission or storage of the data.
If any digit in the ISBN is changed, then the check digit will be different. This means that the check digit can be used as a way of verifying if the number has been altered. (If you want to find out more about ISBNs, check out the International ISBN Agency's website.

Calculating checksums is a lot more complicated than ISBNs, but the principle is the same. They work by performing actions on the contents of the digital files which produce a string of characters unique to that item. If any part of the digital file is altered and the actions are performed on it again, the resulting string of characters will be different. This indicates that the file has changed. A number of different algorithms are available, with different levels of security, to compute checksums: because they require the computer to read through the whole file, for large collections of files this can be a time-consuming and resource-intensive exercise for large digital collections.

Checksums are used in digital preservation as a check that the file is authentic; that is, it has not been altered. You will know from module 1.4 that authenticity is an important characteristic of the preserved material that we want to maintain into the future, and in many situations (such as legal or administrative ones) it is vital to be able to demonstrate that the file is authentic. Checksums assist us in doing this.

**Activity**

Observe the power of checksums. Create a small text file with a text editor (e.g., Notepad on Windows or Text Edit on Mac) and type a simple sentence. Something like:

The quick brown fox jumps over the lazy dog.

Save the file as test.txt then use [http://onlinemd5.com/](http://onlinemd5.com/) to calculate its checksum (you don't need to change anything on this page and you can ignore all the complexities of this tool - just use the default settings and browse to, or drag and drop your text file).

You will get a long string of numbers and letters that will look something like this: E4D909C290D0FB1CA068FFADDFF22CBD0 (using my own test.txt file). This string is the md5 checksum for the test.txt file.

Now make a very minor change to the file and save it again, perhaps remove the full stop: The quick brown fox jumps over the lazy dog

And recalculate the checksum: 9E107D9D372BB6826BD81D3542A419D6 (again removing the fullstop from my own test.txt file).

Note the very big difference in the checksum in response to a minor change in the file's contents. Experiment with adding or removing spaces or characters and observe the changes in the checksum. Image files are even more interesting to look at. You can try altering an image so that just a single pixel is changed in such a way that the change is invisible to the human eye. A checksum check will immediately indicate that a change has been made.
This should make clear why checksums are such an important part of the digital preservation process.

Watch
This Youtube clip, The checksum and digital preservation of oral history (2012). A simple explanation and illustration of its use in practice.

Conclusion
To summarise this topic:

**Preservation metadata:**
Identifies the material for which a preservation program has responsibility.

- Communicates what is needed to maintain and protect data.
- Communicates what is needed to re-present the intended object (or its defined essential elements) to a user when needed, regardless of changes in storage and access technologies.
- Records the history and the effects of what happens to the object.
- Documents the identity and integrity of the object as a basis for authenticity.
- Allows a user and the preservation programme to understand context of the object in storage and in use (UNESCO, 2003, p.94).

Preservation metadata schemes continue to develop, and much more research and evaluation is needed. For example, we still don't know much about how metadata will assist in the effective management of preserved digital objects over time – we are sure that it is essential but have very limited experience on which to evaluate what constitutes the most effective approach. In addition, we need to develop metadata extraction tools and we need to know a lot more about the minimum amount of metadata that we will need for digital preservation to be reliable and effective.

**References**


Module 3.2: Storing digital resources

Introduction

An institution’s digital repository (which may have a label such as ‘library’ or ‘archive’) may include terabytes of data spread across thousands of objects in hundreds of formats. It may, for ease of handling and control, want to transfer the contents to a more readily managed medium. The sheer ease of producing and proliferating electronic files, combined with the perception that storage is so cheap as not to be a concern at all, leads to an uncontrolled explosion in numbers and storage volumes (Linden et al., 2005).

This quote, from a report prepared for the British Library, points to some of the major issues that complicate the issue of digital storage. Archiving such resources for the long term to ensure their ongoing security and accessibility requires substantially more planning and forethought than is necessary for the physical collections held by libraries and archives. The literature talks of ‘sustaining’ digital resources with the need to establish clearly defined policies, procedures and practices to ensure that material is stored in a manner which provides appropriate levels of security while also enabling effective, controlled access.

In terms of general storage approaches, digital resources may be stored either online, offline or nearline, with institutions usually adopting a combination of these approaches. The decision as to which approach or approaches to take is guided primarily by the level of access required but also involves resourcing constraints, consideration of security levels and the longevity required of the digital objects.

**Online storage** refers to materials stored on file servers or other hardware which is immediately accessible to the enquirer via a desktop PC or mobile device. Such an approach provides instant access to these digital objects which is attractive if they are used regularly. However, it does raise issues related to the amount and cost of online storage necessary, particularly as large volumes of digital materials are acquired, response times and the related architecture of very large file structures and their indexes, and concerns over the security processes necessary to ensure such readily accessible files retain their integrity.

**Offline storage** refers to materials that are stored on devices that are not continuously connected to the computer network. Thus data may be stored on magnetic tapes, DVD or CDs, removable hard disk drives and so on, which once created can be removed from the network and stored either locally for easy availability or off site in secure storage for long term safety. Some cloud storage offerings (see below) are essentially remote “offline” storage.
Nearline storage refers to materials stored on removable devices but still connected to a network via some form of hardware and software device. Typically, these are described as a form of 'jukebox', usually comprising magnetic tapes or CDs, where an automated system will select the desired item and insert it into a 'player' to provide access. The process is automated but can take a short while, maybe 30 seconds, to load and make the appropriate item available.

In the last few years, storage "in the cloud" has increasingly become an option for more than just backing up your personal information. All "cloud storage" really means is remotely located storage, accessible over a network. Companies such as Apple and Google offer free storage, sufficient for personal files, or you can subscribe to get access to significant space - terabytes worth from some providers. In fact, "cloud" storage could adopt any or all of the three general approaches described above — Amazon's Glacier seems to be offline storage (although this is never stated explicitly), since Amazon specifies that data in Glacier has a three hour retrieval delay — so is not really an approach but more of an architectural or design choice. Commercial and government organisations are looking closely at the option - there are potential cost savings involved that can be very attractive. However, there are also considerable risks involved for organisations concerned with long term preservation - after all, you are placing your data into the hands of another organisation and trusting them to do the right thing. This might be fine for your personal files and applications...but is it appropriate for say, storing the photographic collection of a museum or national library? As a storage option the 'cloud' needs careful consideration but may well be a viable option.

The choice of which approach to adopt in your organisation very much depends upon the nature of the records and the level of access required. In addition, there are issues such as security, the complexity of the records, the retention span envisaged (five years or ‘forever’?) and the value of the records.

For all digital objects, wherever stored, there is need for establishing audit procedures to monitor who accesses a record and any changes that may have been made to it, even if by an authorised person. In addition, there needs to be some form of automated process that can be run regularly to undertake checking on all objects to ensure they are still accessible, that there has been no change due to drop out or deliberate manipulation, and that indexes and other pointers are still functioning properly. Such a checking program will report any instances of problems with the data.

Note that this maintenance process is particularly important for material stored offline which, by its very nature, is likely to be rarely accessed (and thus problems may go unnoticed for a very long time).

Read
For some of the pros and cons of the various storage strategies, refer to the summary advice developed by the National Archives of Australia.
Consider how an organisation known to you maintains access to their digital material. Is it all available online? Or is offline storage used? Are there any checking procedures in place?

Read this short piece - CNet Australia (2012) comparing cloud services: OneDrive, Dropbox, Google Drive and Box: which cloud storage service is right for you?
and consider this report from the Cloud Security Alliance (2013) on the top nine cloud computing threats.
If you are interested there is a longer report, from Audiovisual Preservation Solutions (2014), Feet on the ground: A practical approach to the cloud. A valuable report providing some good information as to the factors to consider when thinking of using the 'cloud'.
Finally, the Minnesota Historical Society commissioned a study in 2013 on digital preservation and the cloud, which is useful reading and still relevant.
Storage media

As discussed in module 1.2, all digital storage media have a limited life-span that is much shorter than the traditional storage medium for information – paper and ink. The main factors that affect the life expectancy of any digital media are:

- its initial manufacture – can quality be guaranteed?
- The level of access – the less activity the better and;
- the storage conditions in which the media has been kept – cool, dry and dark conditions will provide the most favourable environment.

In addition, the nature of the hardware and software used to support access to the media has to be taken into account. Thus, there is little point maintaining your DVD copies in pristine conditions, aimed at ensuring their longevity for a hundred years or more, when it is likely that DVD players as we know them will no longer be available in 50-years time, and that in the creation of the DVDs a proprietary form of software was used that has not been supported for 20 years or more and is no longer usable. An example of this, familiar to many of you, is the 5¼ inch and 3½ inch discs – one completely obsolete, the other rapidly going that way. Yet you may hold discs with data on them that are in perfectly good condition – but finding a working drive is a different matter. And if the disks were created using say Wordstar or Word Perfect, popular word-processing packages of the 1980s, it will require some form of conversion program to read them, even if you do manage to find a working drive of the right format.

The above example illustrates well that there is unlikely to ever be a really long-term solution to the problems of storage media and their potential obsolescence, and that transfer of files (refreshing - discussed elsewhere in this subject) will have to be a regular procedure (maybe every 10 years or less) for any archive. Obsolescence is no theoretical issue but a major practical concern that has to be taken into account when considering the media most appropriate for your archive in achieving its long-term goals and objectives. One simple technique that all digital preservation systems should adopt is making multiple copies of the objects to reduce the risk of catastrophic data loss. This is a simple technique but one that has obvious cost ramifications.

Storage media generally can be grouped under two broad headings – magnetic media such as hard disks and tapes, and optical media such as CDs and DVDs. Magnetic media rely on the polarity of bits written to the tape or disk which can be either negative or positive. One of the concerns associated with magnetic media is the ease with which this data can be changed – deliberately or accidentally. Optical media use lasers to read and write data and were initially viewed as being more stable than magnetic media due to the possibility of manufacturing them as ‘write once' technologies. This differentiation is no longer so clear and optical media are seen as having problems regarding long term archivability due to the complexity of their makeup (the varying layers comprising the disk, the inks or dyes used in writing information and so on).

Removable hard disks provide high quality storage, but for large volumes of archived data, may be expensive. Digital tapes in the form of cartridges provide high storage capacity at modest cost. Both of these approaches have been in wide use for many years and thus the technology and supporting software have reached a level of maturity that suggests these are as reliable as it is
possible to get for medium-term storage of archival material. CDs and DVDs while cheap and plentiful (at the moment) have generally not been designed for archival storage and should be treated with caution as major storage medium. One of the attractive features of some optical media, however, is its inherent ‘write once’ functionality that, even if not fool proof as noted above, does provide an additional level of security. The difficulty or even inability to alter or remove files from the disc make it an attractive option from the security viewpoint, unlike other storage media that rely upon software or over-ridable hardware controls to limit write access to the media. Recent advances in ultra high density optical storage may offer potential as an archival storage medium combining this security benefit with the reliability of a product developed specifically for high end data storage.

In a digital preservation guide produced by the UK National Archives (2008), the authors discussed a number of criteria that need to be considered in determining the best approach to adopt. They suggest that using a weighting of 1 (does not measure up) to 3 (fully meets criterion) is a useful way in which these criteria can be used to assess the appropriateness of any potential storage media. The criteria to be considered are:

**Longevity** – the media should have a proven life (through laboratory testing and ageing analysis) of 10 years or greater. Considerably longer periods are not seen as a great advantage given the relatively short life expectancy of the supporting hardware.

**Capacity** - sufficient for the needs of the institution in order to minimise as far as possible the number of physical items needed for storage. Taking this approach effectively reduces administrative and management costs and effort.

**Viability** – the media and/or its supporting hardware should include sufficient error detection and correction functionality to ensure the integrity of the data stored. It should also provide an effective and efficient data recovery process to restore lost or damaged files. In addition, media should be either write once (WORM – write once, read many technology) or sufficiently protected to ensure that ideally, damage or loss cannot occur through accident or design.

**Obsolescence** – the media and its associated technology should be in a mature phase – tried and tested with a good size user base, rather than leading edge technology. Those media using open standards should be preferred to those based on proprietary products.

**Cost** – costs must be considered more broadly than just initial purchase price. Ongoing support or maintenance contract costs need to be figured in as well as making some kind of assessment of the costs that may be incurred in the event of failure of the media. Thus Mean Time Before Failure figures (a commonly used IT measure) need to be considered in establishing the overall cost per gigabyte of storage.

**Susceptibility** – the media should be physically robust, capable of withstanding a level of adverse environmental conditions eg. if the air conditioning breaks down on a summer’s weekend.

Using a set of criteria such as these helps in determining the appropriate storage media to be used. However, whichever storage system is adopted, formal backup arrangements must be in place and properly undertaken. In addition, regular checking of the archive files to ensure they are suffering no problems in store is a necessary maintenance routine that has to be built into the workflows for any digital repository. No media can be relied upon for the long term, and again management processes need to be in place to ensure the eventual migration of data from its existing storage platform to a replacement platform at regular intervals (possibly every decade).
Read
For a short overview of storage technologies and future trends read:

Backblaze (2015), Data Storage Technologies of the Future. (NB Backblaze is a US cloud storage service provider)
The Economist (2013). Test-tube data: Archives could last for thousands of years when stored in DNA instead of magnetic tapes and hard drives.
If you are interested in something more detailed, here is a report produced by SpectraLogic (a US data storage company) in June 2017: Digital Data Storage Outlook 2017.

DNA Storage

A very interesting new storage technology, mentioned in the box above, is the use of DNA to store digital bits. DNA storage involves the storage of digital data in the 'base sequence of DNA' (DNA digital data storage, Wikipedia, 2018). DNA storage offers very high storage densities and seems to be a very stable storage medium. It is still early days in the development of DNA storage and there are problems to be overcome but DNA storage seems to offer an exciting new advance in data storage approaches.

If you are interested in reading more here are some resources that can provide you with a little more information about this technology. Remember this is an interesting emerging technology so there are probably 100s more resources on the Internet. If you wish to pursue this topic you can do your own internet and library searches on DNA storage.


Discussion forum
Select one of the newer technologies discussed in this section and research it further on the web. Post comments regarding what you have found to the forum.

RAID

RAID stands for Redundant Array of Independent Disks and is a form of hard disk storage commonly used for storing large amounts of data (terabytes or more). Used with file servers or other larger computers (although gradually moving into use in home PCs), a RAID comprises a series of disk drives linked together with the objective of providing fast access to data and a particularly secure environment for data storage if the higher level of RAID technology is used. The system sees the RAID array as one virtual device and improved access times are achieved
using software that is able to access two parts of the same object at the same time from separate discs, bringing it together for display.

Security is provided by using the mirror imaging possibilities of the RAID technologies where data is written to one disk in the array then duplicated on the next. If something happens to the first disk, the data should be available from the second. In such a situation, the damaged first disk can be replaced without having to bring down the system as data can continue to be accessed from the duplicate disk. This security and data replication mechanism, is why RAID storage is one approach for archival storage that is sued by more organisations managing a digital archive. Even at small scales RAID is an affordable approach to digital storage.

RAID arrays are generally available in number of increasingly sophisticated levels, referred to by numbers, such as RAID 0 and RAID 5, etc. A basic RAID array (RAID 0) will provide a level of mirroring but little else. Further up the scale (RAID 10) sophisticated error checking processes are available, auditing the data by checking what was originally written to what has been accessed and reporting any data discrepancies that may have occurred.

The strength of RAID storage systems stems from the fundamental idea that all hardware is guaranteed to fail at some point in time. A properly set up RAID system can tolerate the failure of several individual storage devices without a loss of data. An operator simply replaces the failed devices and the system rebuilds the data onto the new devices without loss.

Managing the RAID array can either be done through the hardware itself or software or a combination of both. There are advantages in using the hardware approach in that a problem at the operating system level, for example, failure to boot the system, will not make the disk array inaccessible as it does not rely on the software to manage its functions. Software managing the RAID may be faster and frequently a hybrid approach is adopted.

Many larger institutions have adopted a SAN (Storage Area Network) that can manage multiple servers and multiple storage facilities which communicate with each other on a network. Thus the SAN can manage a number of file servers and a number of RAID arrays, optimising the use of storage and access speeds. This approach enables servers and storage devices to be independent of each other so that additional storage can be added or removed with less disruption to the network than would otherwise be the case.

In an environment concerned with digital preservation for the long term, the use of higher level RAID arrays is seen as a good option to ensure that, as far as possible, data will be safe and secure. Supplying this level of confidence in the stored data comes at a cost though, with the sophisticated RAID storage mechanisms being relatively expensive pieces of hardware. While the RAID approach does provide a level of security, the usual backup processes undertaken for any computer system should still be firmly in place for the digital store, as well as the redundant copies as mentioned in an earlier Module. Thus daily backups, weekly backups and storage offsite of the backed up versions are still essential in order to provide that base level backup that may be required in the event of problems. In the ideal world, a complete mirrored system would exist, maintained in a different physical location, as a fall back option should disaster strike!

Read
An excellent and easy to read overview of the various RAID technologies from Advanced Computer & Network Corporation's RAID tutorial. Make sure you check out the various RAID levels and note that some are not really commercially viable.

... and this thought provoking advice from National Archives of Australia regarding the outsourcing of storage generally, Outsourcing digital data storage.
Repositories

Institutional repositories

Institutional repositories have been established in recent years in order to provide an archive for the digital output of an organisation, usually a university or research organisation. Such an archive exists to bring together the inputs and outputs from research, theses and publications, computer programs, art works, teaching materials, course and subject content and information and any related administrative materials that are in digital form; the aim being to preserve this intellectual output of the institution and make it accessible in a standardised form. The desire for accessibility and thus the ultimate goal of enhancing scholarly communication, has led to the development of standards to ensure high levels of interoperability between repositories. If unfamiliar with the repository concept, you might like to check out the ANU Digital collections or the Australian Data Archive.

The move to using institutional repositories to disseminate scholarly information could ultimately have an impact on the traditional methods of journal publication for distributing research knowledge. While nothing radical is likely to occur in the short term, longer term it may be that, as Johnson (2002) has noted:

> digital publishing and networking technologies harnessed by an increasingly dissatisfied library market – as well as by authors themselves – are now driving fundamental changes to this publishing model at an accelerating pace.

Essentially, authors such as Johnson and others foresee the possibility that institutional repositories may come to be the preferred site for the 'official' publication of research papers and that the use of traditional journal publishing for this purpose may decline in the future. There are certainly considerable cost advantages for libraries and their institutions (via a reduction in serial subscriptions) if this approach does emerge. In addition, scholarly communication is made easier through the ready availability of the papers residing in the repository (no need for a subscription) and the common interfaces being adopted in their design. Another driver pushing academic institutions, and authors and researchers to move away from the traditional publishing model is the ongoing dispute about intellectual property rights, which most traditional publishing houses require authors to transfer to them.

There are several established software platforms available which are being used by various institutions for designing and implementing their repositories. Essentially, this software needs to be able to manage the acquisition of materials to determine who can place materials in the repository and the nature/extent of associated metadata required. It also needs to control the format of the material being stored, and avoid proprietary software in favour of commonly supported open source approaches. In addition, the software also needs to be able to provide various levels of rights management and access (for both internal and external users), maintain audit trails and have practices and procedures in place to ensure the long term viability of the archive.

DSpace is an example of freely available institutional repository software, jointly developed by MIT and Hewlett Packard specifically for digital archiving. Since 2002, more than a hundred universities around the world have implemented DSpace repositories, including MIT in the US, Cambridge in the UK, the University of Toronto in Canada and the Australian National University. An active user community has seen the rapid development of the software and its implementation in a wide range of repositories, not confined to universities alone.
Fedora is another such package, jointly developed by Cornell University and the University of Virginia. Again freely available, Fedora has been adopted by major universities such as Cornell in the US, Oxford in the UK, the University of Athens and the University of Queensland. Currently with a smaller user base than DSpace, it is becoming more widely adopted as its user base grows and diversifies. It is a more complex application than DSpace, but that also means it has more functionality than DSpace.

There is some additional discussion of repository software in module 3.4, in the second tools subsection.

<table>
<thead>
<tr>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read these resources to get a better understanding of institutional repositories and their implementation.</td>
</tr>
</tbody>
</table>


and

Open Access Scholarly Information Network (2009). *Establishing a repository*. This online article provides a good overview of what needs to be considered when setting up an institutional repository.

In 2014 UNESCO published a very useful comparison of institutional repository software. The report, *Institutional Repository Software Comparison*, can be downloaded from the UNESCO website.

**Trusted digital repository concept**

The types of activities undertaken by digital repositories, aimed at keeping digital content alive (accessible & useable) over time, have traditionally been the preserve of cultural heritage institutions, like libraries and archives. In the digital age, many different types of institutions, such as universities, private research bodies, government bodies that are creators of significant amounts of digital data, (in meteorology, land surveying and mapping, oceanic research, etc.) and so on, are likely to establish digital repositories to maintain their own outputs and resources. The public trust which is generally afforded to existing cultural heritage institutions is unlikely to be so easily or readily extended to the new players in digital data preservation who will appear in the future.

In 2002, the Research Libraries Group (RLG) and OCLC in the US published a wide-ranging report looking at the whole concept of trust in digital repositories. Such repositories have to be able to manage the identity, integrity and quality of the archived material to ensure users can be confident that the archive will exist unchanged into the future, with records remaining true to their original content (as do existing cultural heritage institutions). The RLG/OCLC report took the view that any new repository had to be developed in compliance with the Open Archival Information System (OAIS) standard (ISO 14721) as its basis, since this has become a *de facto* standard for setting up archival preservation systems. In addition, the report discussed other factors that need to be considered in building a responsible repository including:

- Auditability
- Security
- Backup and recovery systems
Agreements with the creators and providers of data

The report also outlined a framework that would have to be in place to ensure that the repository was in fact one that could be trusted to meet its long term goals. This framework talked of:

**Administrative responsibility** – relating to adopting national or international standards, involving users and the broader community in developing appropriate policies and processes and being open and transparent in all its dealings.

**Organisational viability** – that the organisation has appropriate resources and commitment to ongoing support of the archive and its development.

**Financial sustainability** – good business planning should be in place and proper accounting standards and financial statements made available. Evidence of long-term planning and financial commitment is required.

**Technological suitability** – establishing appropriate technology management strategies to ensure current and ongoing acquisition and support of appropriate hardware and software technologies. Policies and plans need to be in place to monitor technological change, and signal the need for upgrading or replacement of equipment.

**System security** – policies, plans and processes must be in place to address disaster situations as well as detect problems with lost or damaged data, and provide appropriate backup and recovery mechanisms. Audit trails and related security procedures will be integral to the system.

**Procedural accountability** – all processes will be well documented and maintained and made openly available.

To ensure the adoption of a framework such as this and the relevant standards of practice, the report noted that self-assessment may be insufficient and recommended some form of certification on a regular basis by a relevant national body. This was backed up by the publication of a detailed report in 2005 developed by RLG and NARA (the National Archives and Records Administration in the US), *Attributes of a trusted digital repository*, which provided more detail on an audit tool to be used for the certification process. This report provided a considerable level of detail in the form of checklists and report cards aimed at making it a practical tool for any digital archive wishing to undertake a detailed level of assessment. At the time of writing, a small but growing number of institutions are adopting the trusted digital repository model and using checklists such as this to ensure their compliance.

That report provided a foundation upon which the development of trusted digital repositories standards has progressed. A working group set up following the publication of the report led to the development of criteria for a trusted digital repository. This was published in 2007 as *Trustworthy Repositories Audit and Certification*. Continuing work by this group and demonstrated interest from the library and archives communities eventually resulted in the publication of ISO 16363:2012 *Audit and certification of trustworthy digital repositories*. This standard "defines a recommended practice for assessing the trustworthiness of digital repositories. It is applicable to the entire range of digital repositories". ISO 16363:2012 can be used as a basis for certification and has led to ongoing work to develop a standard for certifying bodies who can undertake audit and certification to the ISO 16363 standard. This work resulted in the publication of ISO 16919:2014 *Requirements for bodies providing audit and certification of candidate trustworthy digital repositories*. 
Activity
Investigate this Centre for Research Libraries page for links to relevant documents and other information about trusted digital repositories. Browse some of the linked pages/documents.

Read
For a view from Europe, read the following article about Nestor, a German initiative. The criteria outlined here are similar to the RLG/NARA criteria and are clearly presented in this paper, The nestor Catalogue of Criteria for Trusted Digital Repository Evaluation and Certification (2007). This 2015 article from Bernadette Houghton of Geelong University discussing a self assessment against the ISO standard: Trustworthiness: Self-assessment of an Institutional Repository against ISO 16363-2012. D-Lib Magazine 21(3/4).

References


Module 3.3: Managing digital preservation

Introduction

In the introduction to this subject, we considered why preserving information was important. Let’s consider it a little further. We have already noted the cultural drivers behind it, but also consider the role of records more generally in our society. For individuals, your birth certificate, taxation records, health records, etc. are basic defining records about you. For business and government, proper and accurate record-keeping is fundamental to a modern society. And because virtually all such records are now created electronically, there is a requirement for effective policies and practices to ensure records are created, managed, preserved and made accessible. A commonly used term for this whole digital process is digital continuity, of which digital preservation is a key part.

Digital continuity is "the ability to use digital information in the way that you need, for as long as you need" (The National Archives (UK)). It is all about ensuring that digital records are preserved over the long term according to legislative or other requirements in a manner that ensures their authenticity (i.e., you can be sure that their content is true to the original file even if it has been copied or migrated over time) and ready accessibility. The UK National Archives summarise this by noting that digital continuity is about information usability which comprises the ability to:

- Find the information when you need it
- Open it when you need it
- Work with it in the way you need to
- Understand what it is and what it is about and
- Trust that it is what it says it is. (NB This is not about trusting content)

This is particularly important for organisations – government, business, etc. – which have a requirement to maintain records over the long term to provide evidence of actions, and for accountability. The first two readings below illustrate what can happen if organisations do not have proper procedures in place.

Thus, organisations of any size are advised to develop the necessary strategies and infrastructure to help establish an environment where policies and practices related to the management of digital information over time are put in place, and the activity is properly resourced.

Successful digital preservation requires a mix of technical, managerial and cultural understanding in order to develop policies, procedures and systems that will ensure that the best use is made of the organisation’s scarce resources to achieve its digital preservation goals. The responsibility for pulling together information from this range of stakeholders, balancing it with other organisational priorities and using this to develop both long-term strategies and short-term plans that take into account industry best practice has to lie with the senior management group within the organisation. JISC (Joint Information Systems Committee, a UK body concerned with ongoing access to scholarly and educational resources) put out a wide ranging ‘briefing paper’ on digital preservation which included the following comment regarding management responsibilities:

Resource managers, who may also be resource creators, are responsible for ensuring information is properly managed and remains accessible while in their care. Insofar as preservation is concerned, this can involve everything from preservation planning for avoiding obsolescence, to developing strategy, taking responsibility for preservation, and liaising with IT staff or external preservation service providers (including national bodies) who provide the technical tools and infrastructure for digital preservation (JISC, 2006).

The importance of good management for any complex project (and digital preservation is, as you have no doubt realised during this course, a complex business) is undeniable. Bringing together all the requisite resources, at the right time, in the right sequence, and in the right priority to meet the organisation’s aims and objectives can be a demanding task. This topic considers some of the
most important of these tasks and how they may be approached in order to help ensure the success of the project.

Read
This 2012 press release from the UK Financial Services Authority: Bank of Scotland fined £4.2 Million. Establishing a digital preservation policy - useful advice about developing digital preservation policies from JISC Digital Media.

Listen
To Bob Pymm discussing management issues in relation to digital preservation.

**Policy and priority development**

Creating and implementing some form of digital preservation plan has become a key strategy for a wide range of governments and industries. Such a plan aims to ensure that, as well as having appropriate policies and practices in place, responsibilities are clearly defined and assigned. After all, it is a complex decision in any large organisation as to who does have this responsibility – is it the creator of the record or someone else? The Records Management section or individual? The IT department? Or other corporate entity? Unless tasks are clearly allocated, there are risks involved with information being lost or corrupted in some way. As we have already noted, digital information is a fragile entity and cannot, like paper, be left to sit quietly in the background until needed, say in 50 years time. And as those earlier readings noted, there is a real cost in not having a plan or strategy in place to ensure the availability of digital information when it is required. It needs active management to ensure its availability – and accessibility – in 2060.

The National Archives of Australia have published Digital continuity principles (2011) to inform the development of a continuity (and any preservation) plan. These principles are:

- Digital information is valued as an asset to the organisation and should be managed as such. The costs of poor digital information management need to be explicitly understood. There needs to be recognition that value may change over time and that such value will be different for different stakeholders. Thus priorities need to be determined in the light of these considerations.

- Digital information management is a natural part of overall organisational governance. Clear roles and responsibilities need to be assigned with support and resourcing from senior management. It is an organisation wide responsibility and proper coordination and communication across all areas is vital for the success of any plan.

- The reliability of digital information is key to its usefulness. Trust in the information must exist and appropriate controls, security and audit practices need to be established in order to stop unauthorised access and record all access to the data.

- Having the ability to easily and quickly discover and then access the appropriate information will ensure that the digital information can be readily evaluated and understood, shared and repackaged as needed.

- Digital information should be managed digitally, without the need to output to a physical medium which increasingly, is not possible given the nature of the original digital files and the possible loss of context and content transfer to a physical medium will cause.
Digital information needs to be managed for as long as it has a useful life. This is not always predictable, though appraisal, retention and disposal schedules apply equally to physical and digital media, and care must be taken to actively manage data as it moves beyond the life of the system within which it was created.

These principles should be taken into account when developing both digital continuity and preservation plans. Remember, this process is not something that can be created in a vacuum, but needs input from staff throughout the organisation, including those responsible for providing access to the collections, those responsible for preservation and those involved with building and maintaining a stable network environment. In addition, key external stakeholders may also be involved and their opinions sought.

A digital preservation policy needs to sit within a broader organisational policy framework to ensure coordination and cooperation in realising organisational goals. Thus, the principles behind this policy are no different to those behind any of the other major policies that shape how an organisation goes about achieving its aims. Such principles include:

- **scope** (what is covered in the policy and what it applies to)
- **roles and responsibilities**
- **aims** – what is hoped will be achieved through this policy
- **resourcing** – at what level are staff and funding to be made available
- **how this policy fits in with the broader policy framework**
- **some form of monitoring and evaluation program to consider how successful or otherwise the policy has been.**

The Canadian Heritage Information Network (2013) have developed a [Digital Preservation Toolkit](#) focused on museum requirements but with a wider applicability which provides a number of templates or tools that "offer concrete steps to identify digital material found in your museum, the potential risk and impact of lost material, and how to get started in the development of Preservation Policies, Plans and Procedures". This is a comprehensive resource which is particularly useful for smaller or medium-sized organisations who have limited expertise and resources in this field. Similar initiatives have been undertaken elsewhere in order to try and simplify this seemingly complex process.

Any comprehensive policy will contain a section concerning the prioritisation of materials for preservation (this also applies to conventional preservation policies targeted at analogue materials). Usually, decisions
have to be made as to what will be kept and for how long. In determining what will be kept, criteria common to the analogue collections are also relevant to digital collections. These include:

- A legal responsibility to preserve an item for a fixed period (e.g., government archives, business records and so on).
- The level of uniqueness of an item or group of items – is it available and being preserved elsewhere or is it only held by the one institution?
- The importance of the item or group of items.
- The level of access that will be permitted for an item.
- The role of an item as being representative of a broader group.
- Anticipated use of the material in the future.
- A desire to build a comprehensive collection in this area.

In addition to these basic considerations, additional criteria need to be applied to digital materials. These may include:

- What am I trying to archive – purely the information content of the object (say with emails) or also, the look and feel of the object, for example, a website.
- Ownership and rights concerns – does my institution have the right to preserve these digital objects, or do we have to obtain permission from the owners?
- Is the format in which the item was created one which can be successfully archived or normalised for archiving given the existing system limitations?
- Is sufficient storage available, particularly for very large files (such as moving image data).

All of these considerations need to be taken into account when prioritising material for preservation. They should be clearly articulated in a policy document with corresponding procedure documents that explain how, in practice, this prioritisation process will be implemented. Inevitably, for digital materials, a high level of automation in the prioritisation and subsequent archiving process is necessary in order to manage the volume of material. It is thus common practice to consider groups of materials, rather than each individual item, with set preservation values or flags assigned to each group indicating their priority for ongoing archiving. These flags can be recognised by various archiving software applications in use, triggering (or not) the preservation process.

For digital objects, the prioritisation process raises more concerns than for analogue materials due to the fragility of the digital object. A paper-based item will generally survive quite well if stored in reasonable conditions for hundreds of years with no one doing anything proactively to assist in its preservation. This is not the case with digital objects where, if a decision is made not to actively preserve an item or group of items more or less at the time of its creation (remember the lifecycle concept?), then it becomes increasingly vulnerable to loss through obsolescence within the next 10–20 years. Thus prioritisation for preservation plays a far more important role for digital materials than it does for their analogue counterparts.

Managing costs

Many of the costs involved in digital preservation are the same sorts of costs as would be involved in managing analog records for the long term. However, the heavily technological nature of digital preservation means that there are additional costs, such as software purchasing and maintenance (mentioned above), that have not been associated traditionally with managing analog materials.

Not surprisingly, there has been a considerable amount written about the costs of digital preservation and how these can (or cannot) be justified with regard to the organisation's overall objectives and a number of projects looking at the issue of costs in digital preservation. Before any organisation can be convinced of the justification for committing significant sums of money to the preservation of digital objects, decision makers need to be able to quantify ‘why’ it is necessary to undertake this work and what benefits it brings to the organisation and its purpose. Thus,
developing an understanding of the value of digital assets is a necessary corollary to the work of assessing the costs involved – in the end, the cost/benefit equation has to make sense. This topic was very briefly raised towards the end of module 1.4 and you might like to look at the reading list of that module for a couple of other relevant readings on the cost of digital preservation.

Read
A very good paper discussing this issue of why an organisation might commit to digital preservation is noted below. Consider how well these arguments may apply to your own organisation.


Broadly, the costs of digital preservation can be considered in two parts – those relating to the selection and description of a digital object for preservation and the infrastructure needed to hold and maintain access to that object over time. Note that virtually all of the activities associated with these two cost components require a level of trained and experienced staff involvement, increasing the base costs significantly.

Russell and Weinberger (2000) outline the costs that have to be considered for the first activity of selecting and describing the item for preservation. These cost factors, based on Russell and Weinberger comprise:

- Initial selection of material for preservation. Usually there is a policy and ideally, some form of automated identification of groups or classes of material so that this becomes an automated process with very little in the way of ongoing costs once the software is written.
- Negotiating the right to archive the material and the right to provide access in the future. For an institution such as a government archive this may not be a consideration. For libraries and others dealing with digital objects which they themselves did not create, however, this can be a time-consuming process, requiring a significant commitment of resources. For institutional repositories, it is important to ensure permission to archive and allow access is given as part of the initial deposit process of placing material into the archive.
- What will need to be done in order to archive the object. Does it need to be normalised? Is it in a format easily supported? Does it need to be checked for viruses? Is it just the information being archived or the whole 'look and feel' of the object?
- How will the integrity of the archived object be validated? Will it require human intervention (costly) or can it be done through an automated checking process?
- The completeness of the metadata associated with the object needs to be checked and where necessary created to the appropriate standard.

The second set of costs refers to the technical infrastructure necessary to support ongoing digital archiving activities. These can be considered as:

- Initial capital outlay on hardware and software and the need to budget for updating/replacement of both at regular intervals. This includes the possibilities of running duplicate or 'mirror' systems in order to provide increased reliability, and associated network components.
- The tailoring of software to meet individual archive needs.
- Running costs associated with the system including licence, maintenance and support agreements with suppliers, purchase of new storage media, conducting backups, staff costs associated with system administration, disaster recovery, etc.
- Overheads related to the physical space including air conditioning, lighting and so on.
- Costs involved in the running and checking of audit programs to ensure the integrity of data. Associated with this is the need to maintain a testbed system on which to pilot new or amended software before moving into 'live' status.
- Costs involved in refreshing or migrating the data or developing emulation systems in order to maintain accessibility.
Sanett (2003) brings in a third area relating to the provision of access as another set of costs that needs to be considered in calculating the overall cost of running the digital archive. She notes, ‘It is clear that the soft-funding scenario of the past and present is not sufficient to fund present and projected activities to preserve electronic materials.’

By soft funding, Sanett is referring to the lack of rigorous cost/benefit analysis and comparative costing in the area of digital preservation. She goes on to call for the development of clearer costing models to assist institutions in more reasoned planning for the long-term viability of their institutional repositories.

In the final analysis, it is very difficult to calculate precise figures for the cost of digital preservation given all of the factors, discussed above, that need to be taken into account. However, it is a major factor in the establishment and ongoing maintenance of any digital archive and thus has to be given considerable attention.

**Activity**

Browse these two sites and read the article:

- The website of the Life Project (2005-2010) which "developed a methodology to model the digital lifecycle and calculate the costs of preserving digital information for the next 5, 10 or 20 years."
- This website of the more recent (2013-2015) 4C Project: Collaboration to Clarify the Costs of Curation, which set out to investigate the total cost of curating (= archiving & preserving) digital objects over time.

**Read**

A very good breakdown of the costs of establishing and running an institutional repository are provided in this reading, Kenney, A. (2005). The Cornell experience: arXiv.org. Read this carefully and note the conclusions, including that staffing costs are still the most significant item in the cost of running the archive. What does this suggest will be the future direction for this and similar organisations?

Wheatley, Paul (2012). Digital preservation cost modelling: where did it all go wrong? Emphasises the complexity of estimating costings and makes some suggestions about how the community might progress in this area.

Poole, N., Collections Trust (2010). The cost of digitising Europe’s cultural heritage.

**Risk and business continuity**

As in any enterprise, consideration has to be given to potential risks that may be faced by the archive, the possible severity of those risks and the likelihood of them occurring. No risk can be completely eliminated but with proper policies and associated practices and staff training, their impact may be reduced or avoided all together.

The first step in developing a risk management plan is to identify the risks that may be faced by your institution. Most of these are common to any organisation reliant upon a sophisticated computing environment for the storage and delivery of services. Thus, it is quite appropriate to refer to others’ risk management plans to assist in developing your own approach.
Rosenthal et al. (2005), of Stanford University in the US, have created a list of the likely risks faced by a digital archive which any management plan is going to have to take into account. These comprise:

- **Media failure**: All storage media will degrade with time, leading to minor loss (e.g., a track unreadable or a disk sector corrupted) or catastrophic loss (e.g., a disk crashes and everything on it is lost).
- **Hardware failure**: Hardware of all types will suffer problems over time. Again these can be trivial and relatively easily resolved or major, causing network downtime and major problems for users.
- **Software failure**: Software is usually a highly complex set of instructions that, however well tested, is likely to contain problems or errors that could cause difficulties – a particular problem when using newly released software.
- **Communication errors**: With the large amount of data being transmitted around networks, it is likely that occasionally things will go wrong and data is lost. Rosenthal et al. quote this statistic: ‘A recent study suggests that between one (data) packet in every 16 million packets and one packet in 10 billion packets will have an undetected checksum error’. While this doesn't sound a lot, if you consider the amount of data traveling around the internet at any one time, it does seem more likely that it could happen to you.
- **Failure of network services**: The failure of a network, or any component within it, will pose major difficulties for any organisation reliant upon that network. Thus maintaining network stability is a high priority.
- **Media and hardware obsolescence**: Even if the media or hardware is still running without mishap, as they become obsolete support and maintenance is more difficult to find. Over time, linking these components with more modern systems may prove difficult or impossible.
- **Software obsolescence**: Similarly, software components will become obsolete and files reliant upon this software difficult to access by other systems.
- **Operator error**: Human error will always exist and this can range from trivial incidents affecting one record, to major disasters where entire databases may be irretrievably damaged.
- **Natural disaster**: Depending upon where you are situated, natural disasters such as flood, fire and earthquake must be anticipated.
- **External attack**: While unlikely, this is not unknown and has happened to paper-based libraries in the past. The famous 1986 Los Angeles Public Library fire, was later found to have been started by an arsonist. It is estimated that 45 fires in libraries in the US during the twentieth century were deliberately lit.
- **Internal attack**: Possibly by a disgruntled staff member, which could have a devastating impact.
- **Economic failure**: Resulting in insufficient funding to maintain the archive over the long term.
- **Organisational failure**: Where the parent organisation no longer sees itself in the digital archiving business and wishes to dispose of its asset.

Any organisation, then, needs to take this list of potential risks and assess the likelihood of their occurrence and the impact they may have. This then enables realistic policies and procedures to be developed to minimise the likelihood of the most devastating and likely events from occurring and to manage those less likely, and less damaging, through other approaches.

**Discussion forum**
A common approach to assessing risk and its potential to harm an organisation is through the use of a table or matrix taking into account likelihood, impact and cost of prevention.


Discuss these guidelines and their usefulness on the forum.
Business continuity planning

Business continuity planning refers to the planning and procedures in place to ensure that an organisation can continue with its core business in the face of a significant one-off incident or ongoing problem that will adversely affect its operations. These plans and procedures grow out of the risk management analysis and prioritisation that derive from the risk analysis matrix described above.

For digital archives, loss or even the short term unavailability of its holdings could cause serious disruption to those reliant upon those records. Longer term, the consequences of major loss could be catastrophic. The New South Wales government Information Office has prepared substantial guidelines dealing with business continuity and disaster preparedness. They suggest the following needs to be considered:

- A general policy statement as to how this plan fits into the broader, strategic direction of the archive.
- An up-to-date table identifying staff responsibilities and including contact details for key staff and relevant external bodies.
- A copy of the risk management matrix clearly identifying the most likely potential disasters and their impact.
- The order in which steps should be taken to minimise or control the situation.
- Procedures for assessing a situation and escalating the response as appropriate.
- A listing of any physical items – records or equipment (such as storage media), that may be particularly vulnerable and require specific attention.
- Clearly identified priorities for dealing with the situation.
- Details of equipment and materials available for use in disaster salvage and recovery – what is available, where they are stored and so on.
- Building plans identifying potential site risks and utility points, for example, fuse boxes, water mains taps and so on.
- Provision for staff training and current awareness.
- Regulations regarding emergency purchasing, details of suppliers and insurance cover.

One of the major issues regarding the business continuity plan is to ensure it is maintained and up-to-date (particularly in regard to contact telephone numbers) and that staff are familiar with its contents and understand how to action the plan.

In addition to the comprehensive plan described above, it is useful to establish cards or A4 sheets covering the more likely scenarios which summarise the action required. An example card, based on one developed by Sanders (2004), is shown below

<table>
<thead>
<tr>
<th>Action summary</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Risk</strong></td>
</tr>
<tr>
<td><strong>Probability</strong></td>
</tr>
<tr>
<td><strong>Impact</strong></td>
</tr>
<tr>
<td><strong>Likely scenario</strong></td>
</tr>
<tr>
<td>Functions affected</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
</tbody>
</table>
| **Action**        | • contact Help Desk x xxxx to log problem and establish nature and duration of problem and if necessary request recreation from backup  
|                   | • advise senior management and outline likely impact x xxxx  
|                   | • contact senior manager IT x xxxx to ensure problem is treated with urgency  
|                   | • advise all affected business units  
|                   | • if problem cannot be fixed within one day by recreation from backup, discuss with all stakeholders how to manage current operations and any ad hoc document recreations so that a later full restoration does not make even more problems  
|                   | • if problem cannot be fixed by recreation from backup, investigate ways and need to recreate from paper files, or from individual staff members or suppliers documents etc. – involve all stakeholders  
|                   | • ensure cause of problem is adequately documented and practices put in place to minimise a repeat of this situation. |
| **Responsibility**| Senior manager responsible for document management |
| **Mitigation**    | IT backup – local or offsite. Some data held on local discs in staff PCs |
| **Constraints**   | Time taken to organise backup; potential need to bring down system, making it unavailable for all use |
| **Resources**     | IT and document management staff |

**Activity**
Browse through the Records management disaster planning toolkit, from the State Records Office of South Australia (2007).
This comprehensive document covers the entire subject – risk assessment, planning, response and recovery.
Does your organisation possess such a set of guidelines? Consider the effort required to develop and maintain this document. Also consider the possible costs to an organisation of not producing guidance along these lines.

**References**
4C Project: Collaboration to clarify the costs of curation. Website: [http://4cproject.eu/](http://4cproject.eu/)


Module 3.4: Software tools

Introduction

Regardless of the approach chosen to undertake digital preservation activities, the fact that digital information is mediated by computers gives rise to the necessity for software to be a part of the process. Starting in the late 1990s, projects to develop various types of software for digital preservation commenced around the world. Not all of them were successful, but many persist to this day and have created some very useful software tools.

Two things characterise the majority of the currently available software tools. One is that they have come from libraries, archives or universities, and the other is that they tend to be either open source or free for use. Where software is available under an open source license, it's possible for interested parties to collaborate on the continued development of the code and this leads to the creation of more widely useful tools.

Open Source

The term Open Source is widely used but is often misunderstood. It is sometimes confused with notions of Shareware or Freeware or Public Domain but open source software does not have to be free in the monetary sense. To understand open source, it's necessary to first understand what 'source' means in a software context.

When someone sits down at a computer to write computer software, the writing is generally done in a computer language like Java, C or C++ (to name 3 out of thousands of programming languages) and this writing constitutes Source Code. The source code is then run through some other software on the computer (usually referred to as compiler software) which translates the source code into something a computer can understand. The result is a computer program like a web browser or a word processor.

Open Source software is software that is released along with its source code so that any interested person can study, modify and re-use the code.
In the very early days of computing, before the advent of the personal computer, all software was open source. If you bought a computer and some software, the software always came with its source code in case you needed to make changes to suit your needs. In those days almost anyone using a computer had the technical skills needed to make changes to source code, re-compile it and use it. There was a general acceptance that if you figured out a way to make something work better, you would share your source code with anyone using the same type of computer as you.

These days computers are ubiquitous and most software is sold without its source code and the licenses that accompany most proprietary software products are designed to prevent anyone but the vendor from having access to the source. Today, proportionally fewer computer users have the skills required to understand source code, so for many people the lack of access to the code behind the software they use is not an issue.

For those interested in Digital Preservation, open source software can be very useful. One reason for this is that it eases the task of creating new tools if new work can build on existing code. In addition, since the use of software to manipulate data carries with it the risk of altering the data in unintended ways, the ready availability of source code makes it possible to properly examine what a piece of software is doing. Rather than treating a piece of software as a 'black box', access to source code makes it possible to examine the workings and verify that software does what its authors claim.

Freely available tools

There are many, many examples of digital preservation software readily found via the internet. The projects described in the rest of this module represent just a sample of what is available. The best way to understand the software tools used in digital preservation is to spend some time downloading and using some of them.

Activity
Browse the Open Source Initiative software site to learn about this approach to software development and see a wide range of examples. Check out their blog as well.
Re-visit the COPTR website and look at the tools available to undertake digital preservation work.

Listen
To Michael Carden discussing software tools for digital preservation.

Software (1)

Format characterisation

Digital archives are often concerned with accurately identifying the formats of digital objects in order to make decisions about ingest, preservation and access. This has resulted in several
software projects aimed at developing ways to identify file formats and assess individual files against a format’s specification for compliance or conformance.

The National Archives of the UK started building an online database of formats in 2002. The PRONOM database can be searched online or used via their desktop Digital Record Object and Identification Software (DROID).

The Pronom website explains the importance of format characterisation:

By definition, electronic records are not inherently human-readable - file formats encode information into a form which can only be processed and rendered comprehensible by very specific combinations of hardware and software. The accessibility of that information is therefore highly vulnerable in today's rapidly evolving technological environment. This issue is not solely the concern of digital archivists, but of all those responsible for managing and sustaining access to electronic records over even relatively short timescales.

The Harvard University Library has created the JSTOR/Harvard Object Validation Environment (JHOVE) which contains modules capable of characterising a range of formats including AIFF, WAV, HTML, PDF, GIF, JPEG and TIFF.

A tool from the National Library of New Zealand (NLNZ) is the Metadata Extraction Tool, which we met before in the metadata section of module 3.1. While not strictly intended for format characterisation, the metadata extractor seeks to automate the process of harvesting preservation metadata from digital objects.

In 2009, Harvard University combined their own library's JHove tool with DROID, the NZ metadata extractor and a number of other open source file format identifiers to make the File Information Tool Set (FITS):

FITS is another Java based desktop application that can be downloaded and used on most computers. Its output is very comprehensive but its operation is somewhat slowed by running all of the above tools at once.

Format migration

One widely used approach to digital preservation is to migrate data from one format to another. Some methods convert data to the current version of an older format while others convert data into formats selected for their likely longevity. The former approach is often used where a collection of digital material is accumulated and monitored to attempt to predict the obsolescence of formats. This is exemplified by the National Archives of the UK (TNA) and the National Archives and Records Administration in the United States (NARA). Once a format in the collection is judged to be at risk, a decision can be made on an appropriate format to convert that content to.

In contrast, as noted earlier, conversion to formats selected for longevity (preservation formats) is often done in an automated fashion on ingest to a digital archive. This approach, often referred to as 'normalisation', is employed by the National Archives of Australia (NAA).

The NAA has been involved in the development of open source software for digital preservation since 2002. Their collection of software, which together make up the Digital Preservation Software Platform, includes:

- **Xena** - carries out format identification and conversion into open formats
- **Digital Preservation Recorder** (DPR) - a workflow management application
- **Manifest Maker** - a graphical Python application which takes a file or group of files and creates a plain text manifest list of each item. The manifest includes the file name (including directory structure) as well as a checksum of the file.
- **Checksum Checker** - calculates checksums of stored objects and compares them with entries in a database created by Digital Preservation Recorder.

The Xena software performs the two key tasks of determining file formats and then converting files into appropriate standards based open formats. For example, a Microsoft Word document would be converted into the OpenDocument Format for preservation. A full list of the formats supported by Xena is at the Xena [web site](#).

The Digital Preservation Recorder is a complete digital preservation process workflow application that also creates a preservation audit trail and manages other software components including Xena. The Manifest Maker and Checksum Checker are software tools that support the operation of the DPR.

Each of these software tools is freely downloadable as a desktop application and source code is available to study and modify.

To ease the process of installing all of this software on Windows PCs, the NAA has released an all-in-one [installer package](#) which installs and configures the Digital Preservation Recorder along with its key dependencies: Xena; Manifest Maker; Checksum Checker; Java; Postgres; ClamAV; and OpenOffice.

### Web Archiving

Since its advent in 1989 (so its only 29 years old!) the World Wide Web has grown to become a vital medium for global communication, collaboration, research, publishing and innovation. The web has now “become the platform and interface of choice for virtually every kind of information system” [JISC 2008]. The closing talk at a 2012 web futures conference observed that “the Web is the first medium in history that has no way of being inherently ‘kept’; yet it is where the majority of human creation is now happening.” [Williamson 2012]. The web is, by its very nature, fluid, dynamic and ever-changing. We cannot hope to preserve all of it in any comprehensive way, but individual web archiving initiatives are necessary and can help preserve particular subsets of the evidential, informational and cultural phenomenon that is the World Wide Web.

Web archiving is the process of selecting, collecting, storing, preserving and maintaining the integrity of information from the World Wide Web and making it available and accessible over time. Web archiving poses a number of distinct challenges because of the fluid, dynamic and ever changing nature of the web. Since the inception of the World Wide Web, the volume of digital information it holds and provides access to has increased exponentially, although its not easy to get any firm or reliable figures about the actual size of the web. The [wikipedia entry for the WWW](https://en.wikipedia.org/wiki/World_Wide_Web) says that in 2009 Google discovered 1 trillion unique page URLs, suggesting the web is well over 1 trillion pages by now, but personally, I suspect that no-one really has any idea of the size of the web. In any case, the actual size of the web does not matter for our purposes (which is just to give you some appreciation of the magnitude of the web archiving task), it suffices to say that the web is an enormous human endeavour which is deserving of being preserved at least in part. As Julien Massanès, a prominent and informed writer on web archiving, says:

> The World Wide Web…is a pervasive and ephemeral media where modern culture in a large sense finds a natural form of expression…many aspects of society are happening or reflected on the Internet in general and the Web in particular. Web preservation is for this reason a cultural and historical necessity.

The task of preserving websites often begins with a site harvest initiated by a 'seed list' of URLs. Web harvester tools use the seed list as a starting point and crawl (follow) all the URLs they find to new sites not on the seed list but linked to web sites that are. The best known web archive is the Wayback Machine run by The [Internet Archive](https://archive.org), which has been operating since 1996. As of 2014 the Wayback Machine provided access to 400 billion archived web pages. The Internet Archive uses software called [Heritrix](https://archive.org) to harvest the web for their archive.
There are many other institutional web archiving initiatives underway, attempting to ensure that specific areas of the WWW are captured and preserved. For example, the UK Web Archive archives web pages from the UK domain, while the National Library of Australia's PANDORA web archive captures pages from the AU domain.

One of the most important initiatives is the International Internet Preservation Consortium (IIPC), a global initiative which, although not an actual web archiving initiative, has made many important contributions to the preservation of web sites and web pages. The IIPC have, for example released the Web Curator Tool, described as 'an open-source workflow management application for selective web archiving', and provides other tools to assist with web archiving. The IIPC is a consortium of some 49 members, representing libraries (the majority), archives, and other collecting bodies such as the Internet Archive. The IIPC holds annual assemblies which determine its project focus and research interests year to year.

Databases

The archiving and preservation of databases has always been something of a problematic area, not least because database formats are almost all proprietary and incompatible with each other. As well, there are unresolved theoretical issues regarding the proper archival approach to database archiving, and large problems around selection. Fortunately, the last few years have seen more interest in and development of open source relational database management software such as MySQL and PostgreSQL, which might go some way to addressing the incompatibility problem. Archives and libraries have only recently become more active in this area and there is still a shortage of software available to help in the long term preservation of databases.

One tool, that has actually been around for at least a decade, was developed by the Swiss Federal Archives, who initially designed a data format called SIARD for the preservation of database content. The SFA have made available their SIARD software for free download but the software is not open source. However, it has been adopted by a number of institutions working on the archiving of databases as it is still the only free software tool around that will convert relational databases into a flat preserveable format. Commercial tools that do similar things are very expensive and mostly beyond the reach of archive and library budgets.

Activity
Browse the International Internet Preservation Consortium website to get some idea of the sorts of activities the IIPC is involved in - look particularly at the 'Web Archiving' section. And Gary McGath's file format blog which has been running a series of posts examining various file format identification tools.

Read
This very interesting comparison of SIARD and Chronos (a commercial database archiving tool) undertaken by the Austrian Institute of Technology: Database Preservation Evaluation Report - SIARD vs. CHRONOS
Software (2)

Repository and other digital preservation tools

Repositories (the storage component of an OAIS) are central to any digital preservation solution, as the preserved objects have to be maintained in managed storage. There is a lot of specialised commercial (and proprietary) software available for managing storage repositories, as well as some fine and widely adopted open source offerings, some of which are described below.

Originally developed by MIT and Hewlett Packard, DSpace is an open source digital repository platform that is popular among academic institutions. There is an excellent five minute video explaining DSpace available at the DSpace web site and the software can be freely downloaded.

Similar to DSpace is the Fedora-Commons open source repository platform which has proved popular in the library domain.

In 2009 a project called Duraspase was created to bring together the DSpace and Fedora-Commons work under one umbrella. Both platforms continue to have a separate existence, but Duraspase has extended them and released more open source tools. Duraspase now includes the software platform VIVO, a web-based, open source suite of computer software for managing data about researchers, scientists, and academic faculty members, based on semantic web technologies.

As well as software to manage the actual storage of digital objects, other tools to manage the digital preservation workflow have been and are being developed by a number of initiatives around the world. The COPTR site mentioned earlier in these modules (and in this module, above) lists most of these but some of the more relevant software is described very briefly here.

The National Library of Australia has released their open source Prometheus digital preservation software to manage a digital preservation workflow for the various types of digital material gathered by the library in the normal course of accessioning publications. The software is available via the Prometheus web site.

The Florida Centre for Library Information has released their digital archiving software, Dark Archive In The Sunshine State (DAITSS) under an open source license.

A Canadian company called Artefactual Systems has developed an open source collection of digital archiving tools that they are calling Archivematica. Like Harvard's FITS tool, Archivematica wraps a number of other open source tools into one application. Artefactual are also the developers of the ICA supported ICA- AtoM software for archival descriptions.

A joint open source software project of the Dutch National Library and National Archives is the Dioscuri Emulator application, originally developed in collaboration with the European PLANETS project. The software aims to replicate the operation of an obsolete computer to enable the use of legacy software that may not run on modern computers.

Of course, digital preservation isn't confined to open source software. There are commercial interests looking to make profits in the same domain. One such is the UK software development company Tessella, which has been developing their Safety Deposit Box software for several years with the assistance of many European cultural institutions. This is probably the most widely used commercial digital preservation software tool, having been adopted by a growing number of national and state archives and libraries, including the UK National Archives, FamilySearch and the National Library of Australia. Originally called 'Safety Deposit Box', the software was renamed in 2014 to Preservica, at the same time as Tessella released a cloud option for the software. Preservica is now managed by a separate division of Tessella.
Software tools - Open source


Software (1)


McGath, M. (2015). File formats blog. Retrieved from http://fileformats.wordpress.com/. [The series of posts on file identification has looked at JHOVE (11 June); Exif (8 June); DROID & PRONOM (1 June); file [a unix command line tool] (28 May); and there is an Introduction dated 26 May.]


**Software (2)**


Module 3.5: Big data and preservation

Introduction

The focus of this subject has been digital objects as we traditionally consider them – emails, word processing files, cataloguing records, research data, sound or video recordings and so on. While challenging, their management for long term preservation and access is largely being effectively managed using systems and practices developed over the last 20 years.

However, there are challenges emerging in relation to the actual volume of data now being produced and how it can be managed, including ensuring its accessibility in the long term. This has already been faced with attempts to preserve websites. For almost 20 years now, organisations around the world have been attempting to archive websites using a range of approaches. This includes selective harvesting of targeted websites, an early and very successful example of this is the Australian PANDORA project; a whole-of-domain approach (such as the National Library of Australia acquiring all the ‘.au’ web sites at a point in time) or capturing the entirety of the available web, to some degree being undertaken by the Internet Archive in the US. Preserving web sites brings with it a number of challenges:

- technical in that the files can be highly complex with the mix of video, flash and interactive components using a wide range of software platforms;
- legal as in many jurisdictions web sites are objects created and owned by someone or organisation who holds copyright in them and thus has to approve any copying or capture process; and
- discrimination – which of the millions available should be acquired or should all be captured (as in a whole-of-domain trawl).

The strategies now in place throughout the world point to the resolution of some of these issues and the implementation of a pragmatic approach which is ensuring that some, if not all, of the websites available are being preserved together with their functionality.

Another field which has only recently become an area of concern is in the preservation of feature films, now virtually exclusively created in a digital format. The complexity of shooting a film and integrating its special effects and soundtrack leads to very large datasets, often created in a proprietary format that will make long term preservation very challenging. See for instance this discussion of the making of Girl with the Dragon Tattoo (US version) to give you an idea of the complexity involved. This post on David Bordwell's website suggests that long term preservation was not in the forefront of the Director's planning: 'The Girl with the Dragon Tattoo, for example, was shot with the Red One camera on the company's proprietary format R3D. Other cameras don't use that format. So the footage was converted to other sorts of files for viewing and postproduction work.' (scroll down until you get to the heading 'The preservation of born-digital films is going to be the greatest challenge ever to face archivists' by Margaret Bodde)
Read

Pymm, Bob & Jake Wallis (2009). Archiving the Web: does whole-of-domain archiving = information overload?


Nowak, Arne (2010). Digital cinema technologies from the Archive's perspective. A very good overview of what is involved in creating the massive files associated with today's digital film production.

But what about all the other data out there – the 150,000 hours of Youtube video added daily (NB this statistic is no longer provided by YouTube, however the figure did appear on the statistics webpage in 2015); the 300 million photos uploaded daily to Facebook messages and the 500 million tweets every day? Let alone all those images captured by cameras on street corners, highways, shops, banks etc. And the photos and movies we are taking every day with our mobile cameras (note how the Boston police acquired 10 terabytes of data relating to the Marathon bombing in just a few days). Or the scientific data being generated by massive computers – for instance, the Large Hadron Collider at CERN in Switzerland apparently generates one petabyte of data per second – far more than can be effectively managed by existing computing infrastructure. All of this data has to be somehow managed and some preserved for the longer term – it is THE challenge for digital preservation of the future.

Watch

Brewster Kahle talking about The Internet Archive. It is an astounding achievement and ongoing - and this by a non-profit organisation reliant on donations.

Listen

To the introduction to this module from Bob Pymm, on big data and preservation.

Big data

‘Big data’ is generally raw data, captured as it is created with no structure, and no editing. It is characterised by the three Vs: high volume; high velocity; high variety – which pretty much sums it up. It may have low information density (think of all the street cameras showing nothing except an empty pavement) but this is made up for in sheer volume. Such data is now being seen as offering real opportunities across a wide range of activities – from criminal investigation and national security to health, education, marketing and analysing climate change. Matt Buchanan (2013) notes, in a New Yorker article, “Big data promises to deliver previously unknowable insights, undetectable trends and knowledge hidden in plain sight”. Management of such data should help improve decision-making and service delivery, particularly at the government level. Again, using the street cameras example, evaluating the recordings over time could help traffic planners identify
problem areas before accidents happen or politicians evaluating tweets may be able to more accurately assess trends on particular issues, concerns, etc. There are real opportunities to exploit such data if it can be managed properly. There are of course concerns over the use of big data. Privacy is a big one – if you have your mobile phone switched on it is easy to track your movements or your spending patterns can be analysed based on every credit card transaction you make. Generally there is unease about the amount of information governments and commercial organisations may easily acquire about an individual – and then what they will do with that information. All of these are societal rather than technical issues and our responses to them will evolve over time. From the preservation viewpoint the challenges arise are associated with the capture, selection, indexing and long term storage of such vast amounts of disparate data - and the resources this may require.

**Watch**

This 2013 Youtube clip, [The Challenges of big data: Future risks, governance and trust](https://www.youtube.com/watch?v=example) presented by Bill Buchanan (I assume not related to Matt!) talking about the benefits – and risks – associated with big data.

**Browse**

An example of a big data application which may surprise you, [The archaeology data service](https://www.archaeologydata.org) (the final report of the ADS Big Data Project can be downloaded from a link at the bottom of the page).

Any systems handling such large quantities of data have to be highly automated and capable of 'upscaleing', i.e. handling increasing volumes as time goes by. The components of the SCAPE project (see reading below) look at:

- Content analysis and migration - quality assurance
- Scalable monitoring and control
- Large scale distributed data processing
These are all challenging issues when looking at such huge volumes of data; and the potential for any one organisation to effectively build systems to handle such volumes is limited. It may be that collaborative models will offer the only economically feasible option where the storage and processing of the vast majority of the data is undertaken across linked systems. This concept, known as Grid Computing, obviously requires a high level of cooperation and standards to ensure all systems can 'talk' to each other and users accessing the data do not have to worry about different interfaces. This approach is not dissimilar to Cloud Computing but can be made more secure and by linking systems to provide the necessary processing power essential for handling big data.

**Browse**

The Scaleable Preservation Environments (SCAPE) Project. A pan-European project started in 2011 aiming to develop appropriate open-source software platforms capable of handling these large datasets.

A big data implementation based on grid computing. This is the introduction to a conference on the topic - check out some of the papers. Many are very technical but it gives you an idea of the range of applications using this technology.

**Read**

This short article, Schmidt, Rainer (2013). Collaborative preservation infrastructure: A research challenge which introduces the concept of grid computing.

**Conclusion**

We have covered a wide range of material in this subject and hopefully you have gained some idea of the complexities involved in ensuring our digital outputs are not lost to future generations. The complexities involved, and the volume of data being created make this probably THE challenge for the information industry over coming years. And it is an area in which LIS professionals can play a leading role. Cooperating with the IT industry, as well as the creators of the digital objects themselves, in the design of systems and services should lead to an environment where preservation issues are considered before a digital object is even created. Using commonly accepted standards (ideally open source) across all steps of the data creation, capture and storage processes will help simplify the process and contain costs – which in the end is critical for successful digital preservation. The Blue Ribbon Task Force report (see below) talks of digital preservation being “a form of investment, in the sense of allocating resources now in the hope of achieving benefits later” (p.98) while talking of the need for flexibility in this investment as things inevitably change over time.

**Browse**

This is a wide-ranging report well worth browsing – lots to think about here.

You might also like to browse this report, Charles Beagrie Ltd (2012). *Economic impact evaluation of the Economic and Social Data Service* which provides a range of data on evaluating the benefit of this particular research data service (provided by one of the UK Research Councils). Again, this illustrates how the costs of actually establishing such a service can be assessed against the benefits.

As that last reading indicates, if a case can be made that the benefits of an activity outweigh its costs then it will be a lot easier to put together a convincing business case for the investment of considerable resources into an activity. Thus for a film production company, spending hundreds of millions of dollars on the creation of a feature, investing in a complex digital preservation strategy will make a lot of sense as they are likely to want to be able to access that film over many years in order to fully exploit its earnings potential. But for a government-funded organisation, wishing, say, to archive all the tweets made on a specific day (say your National day), making the case is more difficult - there is no real commercial benefit, but a longer term cultural/historical driver which a cash-strapped government may not see as important. So it can be a challenging process to get together the resources and infrastructure necessary to ensure this gets done. Cultural change (changing mind sets and behaviour) is one of the most difficult but also one of the most crucial aspects of the role of digital preservation specialists. Long term thinking about digital preservation will only, in the long run, happen when the digital preservation community can achieve cultural change across their user communities.

Read

Telecoms, Internet and Broadcast in Africa (2014). *Digitisation to enable accelerated growth in Africa*. An interesting report on the economic benefits for developing economies and the challenges (scroll down until you find the article).

In the end, only the future will know how well we have handled the process, and no doubt there will be gaps as well as successes in what has survived and is accessible. As LIS professionals, getting the digital preservation message to your employers, users and colleagues is one of the most important aspects of your work. I hope this subject has gone a long way to making you feel comfortable with taking on these challenges. Good luck!

Evaluation

Now it's your turn! Do make sure you complete the online evaluation form for this subject – we do take notice of your comments and your feedback helps us to update the subject for coming sessions. Thanks for your attention and interest!