08 Digital preservation strategies

Introduction

A digital preservation strategy is a well-considered and documented approach to the preservation of digital objects. For collecting archivists, the purpose of such a strategy is to ensure that access to the born-digital archives accessioned by a repository can be maintained indefinitely. For repositories responsible for personal digital archives this will probably include material created on long-past, recent, current and future computers and devices. This is a major challenge for digital archivists because the technological landscape evolves so rapidly: new hardware, software and removable media are developed and adopted on a regular basis, and new versions often only have limited backwards compatibility. Obsolescence therefore poses a major threat to the survival of digital records and must be addressed in a preservation strategy.

Digital archivists must also be able to demonstrate to researchers that a preserved digital object is authentic: that no accidental or intentional changes have occurred; and that the stated date of creation or receipt, identity of the author and process that created the object are all verifiable. Preservation strategies are all based on the principle that the preserved digital object should be identical in all essential respects to the digital object which left the creator's computer or other device. This means that it is important to understand what is 'essential' in order to protect those aspects of a record which are essential and to measure the success of preservation interventions.

Numerous approaches, all with slightly different emphases, have been discussed in the digital preservation community. The two principal strategies are migration and emulation, although within each there are further subtle variations on the basic approach. A number of other distinct approaches have also been proposed. This chapter of the Workbook examines currently proposed digital preservation strategies and assesses their usefulness in relation to personal digital archives. It also explores some of the principles which inform each approach and some of the practical steps necessary to implement a preservation strategy.

Degree of preservation

A digital object, such as the kind of typical object found in a personal archive (e.g. a speech written in a word-processed document, a budget in a spreadsheet, or a digital image of a family) is defined in the *PREMIS Data Dictionary* as a discrete unit of information in digital form which is comprised of three different levels:

1. **Bitstream:** in its simplest form, a digital object consists of a bitstream, i.e. an ordered sequence of bits (binary zeros and ones). This binary inscription will usually be stored on a physical medium of some kind. A computer system with the correct combination of hardware or software translates the bitstream into something meaningful.
2. **File:** a named and ordered sequence of bytes known by an operating system. The format of a file is laid out in the format specification, which transforms the file from its binary ones and zeros into something which makes sense to a user, i.e. stipulating the proper encoding, sequence, arrangement, size and internal relationships which enable the construction of a valid file of the relevant type (e.g. gif 1989a file). This level represents the transformation of the input bitstream into output for presentation purposes; the physical medium on which the bitstream is inscribed is therefore of no consequence at this level.
3. **Representation:** denotes the set of files needed for a complete and reasonable rendition of an Intellectual Entity. This is defined by PREMIS as a coherent set of content

that can reasonably be described as a unit; this is essentially a conceptual object, or something that a human can understand as a meaningful unit of information, e.g. a website, a report, a photograph. An Intellectual Entity or conceptual object may have one or more digital representations or encodings; for instance, the text of a politician's speech might be saved as both a Microsoft Word document and a PDF file. The underlying encoding of each will differ considerably, but the textual content of each item is identical. A Representation may also be made up of one or more Files, and it is important that the relationships between such component files are clear.

Simply preserving the bitstream therefore does not guarantee ongoing access to a digital object. The digital object has an existence separate from the medium on which the bitstream is inscribed, and successful preservation is only complete when all the significant properties (see p. 232) are maintained and the digital object can be displayed in a meaningful and understandable form. In some cases, this might mean that only the intellectual content (e.g. the text of a word-processed document) is preserved, and that original formatting and layout is not retained in the preserved object. However, in other cases (notably, but not exclusively, complex objects such as interactive resources) it may be important that the 'look and feel' of the original object, and its functionality, is retained or recreated as part of the preservation process.

This point is usefully illustrated by the definition of digital records propounded by the National Archives of Australia: digital records are viewed as 'performances' – the result of interplay between technology and data. A digital archivist is not primarily interested in preserving the physical bitstream, though without this nothing else is possible. The aim is to capture an acceptable representation of the fleeting and temporary performance of the record on screen as it was originally created, viewed and edited by the individual whose archive is being preserved.

Bitstream preservation

Bitstream preservation is the most basic technical layer of any digital preservation strategy. It could be adopted as the sole layer were the repository to confer responsibility for acquiring file- and representation-level access to future users of the material. This might be acceptable in some circumstances, but is not a reasonable approach to the preservation of personal digital archives, which are likely to be used by readers possessing a wide range of technical knowledge from negligible to expert. Nevertheless, digital archivists must address bitstream preservation as it forms the foundation for all other preservation strategies. Whatever the overall approach to preservation taken by a digital repository, it should be a matter of policy that the bitstream of every archival digital object is preserved in its original form indefinitely. The advantages of doing this are largely obvious and include the following:

- More complex preservation actions can be carried out in the knowledge that, if necessary, the digital archivist can return to the original bitstream; this allows for future developments in preservation techniques.
- In some cases (e.g. obscure or unsupported file formats), preserving the bitstream may be all that is feasible; by doing this, and retaining any documentation associated with the digital object, it may be possible for future archivists to access the digital object and preserve it more fully, even if it is currently unreadable.
- Preserving the bitstream in its original state means that the archive is able to maintain one version of the object which has not been subject to data loss or corruption induced by preservation actions.

Even this most basic level of preservation requires a degree of preservation activity, and the preservation strategy of a digital archive should include provisions for ensuring that unaltered bitstreams are preserved intact over time so that the authenticity of digital objects is not compromised.

Bitstreams must be stored on some kind of physical medium. Ultimately no media is 'archival' and all types will degrade over time; media life expectancy claims are statistical averages based on

accelerated ageing tests which can only provide a rough estimate of longevity. Media technology also evolves quickly. New media types supercede older media and the devices needed to read a particular kind of media are often discontinued far sooner than the media physcially degrades. It is therefore important to have a comprehensive strategy in place for ensuring that the bitstreams of digital objects are stored on suitable media at all times and are checked on a regular basis. The following elements should all form part of a physical preservation strategy for digital media:

Ongoing procedures for regularly refreshing storage media. Refreshing (moving to a newer version of the same storage media, or to different storage media, but with no change to the bitstream) should be carried out at specified times; these should fall within the minimum lifespan of the chosen medium as specified by the manufacturer or by independent sources. Refreshment may also be undertaken in response to an increase in error-rates reported by the storage system. After refreshing (and before the earlier media are securely disposed of), fixity checks should be carried out to ensure that no changes have occurred to the bitstream during the transfer process; all such preservation actions should be well documented, using metadata schemas which provide support for this, such as PREMIS (see p. 80).

- There should also be an ongoing programme for performing fixity or integrity checks, such as checksums and taking restoration actions identified as necessary by such monitoring; this is important for monitoring the stability of bitstreams over time.
- Multiple copies of bitstreams should be maintained to mitigate risk; at least one copy should be kept in a secure onsite location and there should be offsite storage for backups.
- Ideally, copies should be stored on different media types (e.g. magnetic and optical) to reduce dependence on any single storage technology.
- Where the same type of media is used for multiple copies, different brands or batches of media should be used to minimise the risk of faults in particular products or batches of the same product.
- The institutional disaster recovery plan should be comprehensive and take full account of any digital media which forms part of the institution's collections.

There are also some general precautions in relation to storage and handling which can be observed to mitigate the risk of physical degradation, whatever the media employed:

- Media should be stored in a contaminant-free environment.
- Environmental conditions should be stable, and any fluctuations in temperature or humidity should be avoided. For mixed collections the suggested temperature is around 20ºC and relative humidity 40%.
- Media should be stored in closed metal cabinets which are electrically grounded.
- Media should be shelved vertically rather than stacked.
- Media should be handled appropriately: hands should be clean and dry; media should be kept in its case except during use; and labels should be written before being applied to the media.
- Access to the media should be limited to trained staff.
- Exposure to sunlight and UV from lights should be kept to a minimum.
- Media should always be stored in the correct cases – preferably a suitable archival quality case.
- Media should be visually checked for signs of damage on a regular basis.
- Media should be allowed to acclimatise to any new temperature/humidity before using and returned to controlled storage immediately after use.

It is important for digital curators to understand the properties of the various types of storage

media available, because they require different hardware and software equipment for access, and have different storage conditions and preservation requirements. Choosing the most appropriate media can maximise the period between refreshment cycles.

Selecting appropriate preservation media

When selecting appropriate storage media for preservation, The National Archives in the UK recommends taking into consideration:

1. The longevity of the media, which should be at least 10 years.
2. The capacity of the media, which should be appropriate for the quantity of data to be stored, and the physical size of the archival store.
3. The viability of the media, which should have robust error-detection methods for reading and writing data. Media should ideally be write-once.
4. The possible obsolescence of the media and its supporting hardware and software; ideally it should be based on mature technology which is widely available. As with file formats, open standards are preferable to proprietary ones.
5. The cost of the media: comparisons should be made on a price to volume ratio.
6. The susceptibility of the media to physical damage.
7. The media's ability to tolerate a wide range of environmental conditions.

When selecting appropriate storage media for preservation, there are three main types to consider: disk, tape and solid state media.

Optical media

This media type includes:

- The Compact Disc (CD): originally just an audio format, but since the development of the CD-ROM there are also CD types specifically designed to store data accessible by a computer.
- The Digital Video (or Versatile) Disc (DVD).
- Numerous variants on both CD and DVD, e.g. CD-R (Compact Disc-Recordable), DVD-R (Digital Video Disc-Recordable).
- New, high-definition optical disc formats, principally the Blu-ray Disc and High Density or High-Definition DVD (HD DVD).

CDs and DVDs store data in the form of pits within a flat surface; the data is accessed when a special material on the disc is illuminated with a laser diode, and the pits distort the reflected laser light. The discs are comprised of various layers (including a dye layer for recordable media and a reflective layer), and the combination of materials used for these layers can affect stability and longevity.

The National Archives produced a scorecard to measure formats against the seven selection factors listed above. Their conclusion indicates that both CD-R and DVD-R can be considered for long-term preservation; CDs have a slightly higher score – rating particularly highly for longevity, obsolescence, cost and susceptibility. Recent research suggests that CD-Rs which combine a more chemically stable dye layer (e.g. metal-stabilised cyanine) with a reflective layer of gold may have a life span suited to archiving.

Definitive lifespans cannot be determined, but under optimal environmental conditions and with infrequent use, the life expectancy of both a CD and a DVD is predicted to range from approximately two years (at a temperature of 28ºC and relative humidity of 50%) to 75 years (at a temperature of 10ºC and relative humidity of 25%). Media integrity should be monitored by reading a sample of disks periodically and media should be scheduled for refreshment at regular intervals pre-dating the lifespan suggested by the manufacturer. Buying the highest-quality optical storage media can extend the period needed between refreshment cycles.

The new Blue-ray Disc and HD DVD both differ from other optical media in that a blue-violet laser is used for reading and writing data. This has a shorter wavelength than the red laser used by CDs and DVDs, so substantially more data can be stored on a single disc. Currently, Blu-ray provides a single layer storage capacity of 25 GB and HD DVD provides 15 GB (in contrast to the 4.7 GB of conventional single layer DVDs). Technical differences make the two high-definition formats incompatible with each other; this has resulted in a format war comparable to that between Betamax and VHS in the 1980s, with major companies like Sony, Philips and Apple backing Blu-ray, and Toshiba and Microsoft backing HD DVD. In the light of this ongoing competition, and the current lack of knowledge about how well-suited these formats are to long-term preservation, Blu-ray and HD DVD should not yet be used as archival storage media, although they may have great potential for the future and players capable of reading both formats are in development.

Magnetic media

The term magnetic media is used to describe any media format where information is recorded and retrieved in the form of a magnetic signal; the magnetic properties come from metallic materials suspended in a non-magnetic mixture on a substrate or backing material.

The most common types of magnetic media are:

- Magnetic tape: including computer tape stored in cassettes (the open-reel format is now obsolete); and tapes used in digital recording processes, primarily Digital Linear Tape (DLT) and Digital Audio Tape (DAT), which was originally designed for audio use but has now been adopted for general computer data storage; Linear Tape Open (LTO) is a non-proprietary alternative to DLT. The tape consists of a carrier of plastic film coated with a matrix containing magnetisable particles, and a plastic or resin binder, as well as other ingredients.
- Magnetic hard disks, which can be held within a computer, or exist as independent external devices: they consist of a spindle holding one or more rapidly rotating platters which have a metallic (usually aluminium) base coated on both sides with a matrix similar to that of magnetic tape.
- Magnetic floppy disks/diskettes: these consist of a plastic base with a magnetic matrix on one or both sides; this is enclosed in a protective plastic jacket.

Of the magnetic tape varieties, DLT and LTO are high capacity formats. The National Archives scoring system ranks these alongside CD-R, and they are the most stable and long-lasting formats; they are therefore suitable for long-term preservation. So too is DAT, although this has a low storage capacity and scores slightly lower overall than DLT/LTO. Magnetic tape is also the least expensive backup medium per unit. In contrast to optical media, more inexpensive, standard, magnetic tape products can be used for preservation; while this requires more regular and rigorous monitoring and refreshment, it may still be more affordable in the long-term than using the highest quality optical media.

Hard disks have high storage density and are reasonably robust. The kind of hard disks found in personal computers currently hold between approximately 20 GB and 750 GB of data. Larger scale storage employing hard disk technology uses disk arrays, which organise multiple disks into a logical volume of storage. Archival storage should use RAID (Redundant Array of Independent Disks) technology, which, depending on the level of RAID used, can protect against some level of data loss in the event of one or more disks in the array failing.

Hard disk drives have a life expectancy of five years at the most; this means that they would need regular refreshment. Their advantage is that they are spinning disks, and therefore the archive can automate regular fixity checks. The National Archives recommends server-based hard disk storage as the most effective and secure storage regime for electronic records.

Floppy disks have very limited capacity; they are susceptible to accidental erasure and have a very short lifespan. Most modern computers do not include a floppy disk drive, so developments in hardware are rapidly rendering them obsolete. For all of these reasons, floppy disks are not suitable for long-term preservation purposes and any data residing on these should be transferred to more appropriate media as soon as possible.

All electro-magnetic devices are susceptible to electro-magnetic radiation. Electromagnetic Pulse (EMP), which can be generated by nuclear detonations or electromagnetic bombs, is the most damaging form of this. Protection against electro-magnetic effects can be provided by Faraday cages, or repositories could opt to store copies of their data on optical media, although these could not be read until damage to electrical and computing infrastructure is remedied. More common electromagnetic interference is caused by active wireless devices, such as mobile phones.

Solid state media

The term solid state denotes removable storage devices which use flash memory, including USB and memory sticks, and cards which are used in digital cameras or laptops, like CompactFlash, xD Picture Cards, SD Memory cards and MultiMedia Cards. Solid state media holds data in smaller packages than hard drives, which makes their storage more efficient; their capacity does not yet equal that of hard disk drives, but is increasing all the time. They are very small and have no moving parts which make them particularly robust and portable. While these media are useful for short-term portable storage, their archival properties are not yet well understood, and they are therefore not appropriate for long-term preservation use.

Storage and handling of removable digital media

In addition to the general storage and handling recommendations for all types of media, other recommendations can be made for specific media types selected for long-term storage.

Optical:

- For long-term storage, CDs and DVDs should ideally be stored at a temperature of 18-22ºC and at a relative humidity of 35-45%.
- Perhaps surprisingly, the top surface (or label area) of a CD is more vulnerable than the underside and requires extra care.
- Optical discs should only be handled by the extreme edges or the centre hole.
- Cleaning of CDs or DVDs should be done from the outer to the inner edge, rather than along the tracks.
- Optical discs should never be flexed. DVDs are most vulnerable because their tracks are more closely spaced; special DVD carriers can minimise flexing when discs have to be moved.
- The surfaces of optical discs should not be marked, unless according to the manufacturer's recommendation; if marking a disc, use soft tipped pen with water-soluble, permanent ink and only mark the upper surface.

Magnetic tape cartridges:

- Store DLT at a temperature of 18-26ºC and relative humidity of 40-60%.
- Store LTO at a temperature of 16-32ºC and relative humidity of 20-80%.
- Store DAT at a temperature of 5-32ºC and relative humidity of 20-60%.
- Avoid exposure to magnetic fields: these can alter the media and lead to data loss.
- Minimise handling and use and return tapes to their containers directly after use.
- Tape cartridges should never be opened.
- The tape surface should never be touched.
- Label in ink rather than pencil: graphite dust can interfere with the reading of the tape.
- Before use, tapes should be forwarded and rewound fully to equalise tape tentions.
- After writing, the tape should be fully rewound; tapes should never be left in a partly wound state for any length of time.
- When transporting tapes, use enclosures or packaging with a space clearance of 50 mm around the media.
- Tape cartridges should be retensioned at yearly intervals.
- The write-protect switch should be set after writing.

Solid state:

- Media should only be held by the extreme edges.
- Labels should only be applied within the approved label area.

Institutions should follow the recommendations established in British Standard 4783, Storage, *Transportation and Maintenance of Media for use in Data Processing and Information Storage.*

File formats

In order to process the bitstream of a digital object and convert it into something meaningful, it is necessary to know what file format the bits are in; it is the format specification that transforms a bitstream into a particular type of computer file.

Most personal digital archives contain a wide range of file formats. By the time the archive comes into the custody of a digital archivist, many may be obsolete (unreadable by modern software or hardware) or in danger of becoming obsolete. Various factors contribute to format obsolescence, e.g.:

- Whilst some format specifications are independent of specific software (e.g. ASCII), most are tied to particular software; when the software becomes obsolete, so does the format.
- Much software is proprietary, and when software companies introduce new versions of software these do not always support files created in earlier versions. In addition, detailed technical information about file formats is often not publicly available.
- Some formats are simply unpopular and are discontinued due to lack of use, or lack of compatible software.

When an archive is received, the digital archivist must identify all the file formats included in the accession and validate them to see that they conform to the relevant format specification where this is available. Various registries and tools exist to assist digital curators in this task and to support other activities which form part of an institution's preservation strategy.

Format registries

These are third-party services which contain varying levels of information about file formats and, when more fully mature, could serve a number of purposes for digital curators, such as:

- Providing information about file formats which support decisions on suitable preservation formats (see p. 230) to support within a repository.
- Recommending migration (see p. 235) paths from original file formats to newer or preservation-friendly formats.
- Supporting a repository's technology watch function, which includes monitoring formats for risk of obsolescence; registries could also have an alert function, to warn the archivist as soon as risk of obsolescence has been identified.
- Helping archivists to identify and validate file formats received in new accessions of digital archive material.
- Allowing curators to look up the characteristics of a format, for example to identify automatic metadata extraction techniques.

File format registries can:

- Refer to externally-held file format specifications and compatible software.

- Collect and preserve copies of specifications and software.
- Or a combination of the two.

The first is easier and cheaper to administer, but problems may arise when file format specifications cease to be accessible. This is unlikely in the case of formats which have been registered as standards (e.g. Open Document Format ISO/IEC 26300:2006), but is a major issue in relation to proprietary formats, where earlier specifications may not be made available, or preserved at all, by the manufacturer.

File format registries are being developed and maintained by large (often national) archives and libraries with a strength in digital preservation, to provide a useful source of information for the digital preservation community; these have the promise to become well-established resources. Others are provided independently and are often intended for computer programmers rather than those engaged in preservation activities.

Examples of privately-maintained websites on file formats include:

- Wotsit.org contains short descriptions, file extensions and format specifications for hundreds of different formats.
- File Format Encyclopedia contains similar information.
- File Extension Source contains detailed information about file formats and their associated extensions.

Examples of registries with a long-term digital preservation focus:

- Digital Formats for Library of Congress Collections: provides information on the suitability of various digital file formats for long-term preservation, assessing each format against named sustainability factors and content-type specific quality and functionality factors.
- PRONOM is provided by The National Archives in the UK and contains basic information about some file formats and their supporting software.
- Global Digital Format Registry (GDFR) is being developed by Harvard University Library. The registry will collect representation information from centres around the world, which will be made available as a resource for any repository in the world.
- Global Format Registry (GFR): a prototype registry developed at Maryland University, containing file format and application information; only sparsely populated.

Tools

There are various tools (often closely associated with format registries) which enable archivists to identify and validate file formats; other tools assist with preservation actions such as migration. These include:

- Digital Record Object Identification (DROID): a software tool developed by The National Archives to perform automated batch identification of file formats.
- JHOVE: a tool developed by JSTOR and Harvard University Library to allow the automatic identification, validation and characterisation of a range of digital object types.
- FOrmat CUration Service (FOCUS): a prototype tool which will perform identification and validation on submitted files.
- XML Electronic Normalising of Archives (XENA): a tool for converting a range of file formats to XML representations, used in normalisation.
- Conversion and Recommendation of Digital Object Formats (CRiB): an online migration tool, which recommends optimal migration alternatives, undertakes the conversion process, evaluates the outcome of the migration and generates migration reports in appropriate forms for inclusion in preservation metadata records. It currently supports migration paths for a number of image formats, but can be scaled to provide for other formats.

Some of these tools are online services and may be unsuitable for use with closed archival collections.

Selecting file formats for preservation

It is unlikely that a repository will choose only to preserve bitstreams of digital objects, although this may be a necessary last resort in the case of unusual and obscure file formats, or those for which the format specification is unavailable. Any approach beyond the level of bitstream preservation requires a careful consideration of which file formats are most appropriate for preservation purposes in any particular case. Some repositories may keep files in their original formats, others may choose to support only a limited number of formats and others may normalise to one format. More information on these different approaches (all types of migration) is given below (see p. 235).

In order to preserve digital objects properly, digital curators, or agents developing tools on their behalf, require access to detailed technical information about their file formats.

Some proprietary formats have open specifications, meaning that they are largely independent of specific software and are therefore more suited to preservation, e.g. Adobe's PDF. Other owners of proprietary formats (which, unfortunately for archivists, are often the most widely used and popular formats) have no detailed specification for their formats, or restrict access to format specifications to third-party developers who have signed a non-disclosure agreement, allowing products compatible with their formats to be designed without publishing their specifications openly. Published and open formats are always preferable from a preservation perspective: the best options for digital curation and long-term preservation are non-proprietary, open format specifications produced by international standards bodies, such as ISO/IEC 26300:2006, the Open Document Format for Office Applications. Usually numerous organisations have been involved in the development of these standards and they are generally backwards compatible.

Issues to consider when selecting file formats for long-term preservation include:

- Is it defined by an international, national or publicly available standard?
- Is the quality of the specification adequate?
- How widely has the format been adopted as a preservation format?
- Is it backwards compatible?
- Is it independent of any specific hardware or software environment?
- Does it have good metadata support (i.e. metadata providing technical and provenance information which is generated by the creating application, entered manually by the record creator, or a combination of these)?
- Does it have a good range of functionality without being too complex for the purpose?
- Is it easily convertible into other formats (for migration purposes)?
- How well does it retain the formatting and other significant properties of converted digital objects?
- How stable is the format?
- How proven is it in terms of longevity?
- Does it include an error-detection facility?

Arms and Fleishhauer[1] have codified such selection criteria into a decision-support framework for use when considering preservation formats for Library of Congress digital collections. They have identified seven sustainability factors which apply across digital formats for all categories of information and are applicable whichever preservation strategy is selected. These are:

1. **Disclosure:** the degree of access to full specifications and tools for validating technical integrity; open standards are usually more fully documented and more likely to be supported by tools for validation than proprietary formats.
2. **Adoption:** the degree to which the format is already used; if widely used, it is less likely to become obsolete quickly, and commercial tools for migration and emulation are more likely to emerge from the computing industry which archive institutions can purchase.
3. **Transparency:** the degree to which the digital representation is open to direct analysis with basic tools; this is enhanced if textual content employs standard character encodings.

1 Caroline R. Arms and Carl Fleischhauer, 'Digital formats: factors for sustainability, functionality, and quality', paper from the IS&T Archiving 2005 Conference, Washington DC (29 April 2005). URL: <http://memory.loc.gov/ammem/techdocs/digform/Formats_IST05_paper.pdf>

4. **Self-documentation:** it is easier to manage digital objects that contain basic descriptive, technical and administrative metadata.
5. **External dependencies:** the degree to which a format depends on particular hardware, operating system or software for rendering or use.
6. **Impact of patents:** the degree to which digital preservation will be inhibited by patents.
7. **Technical production mechanisms:** the implementation of mechanisms like encryption that might prevent the preservation of content by the digital repository.

They also identify quality and functionality factors which are genre specific, and pertain to the ability of a particular format to represent the significant characteristics required or expected by current and future users of a given content item.

Another approach to selecting preservation formats is provided by OCLC's INFORM Methodology, which measures the preservation durability of digital formats. It compares formats and preservation approaches, and provides a risk management-based means of tracking what might be lost over time if particular preservation actions are taken; the digital archivist can then make decisions about preservation strategy based on this risk assessment.

If a digital repository chooses to convert digital objects to one or more standard formats, there are a number of candidates for consideration; examples for text-based documents include:

Extensible Markup Language (XML): this is not a format, rather a general-purpose markup language for describing the structure and meaning of data. It is an open standard defined by the World Wide Web Consortium and is independent of specific applications. Preserving digital objects which have been created using XML in accordance with a standard DTD or Schema is straightforward. Converting other digital objects to XML is one kind of migration approach; the National Archives of Australia, for example, normalises the formats it receives to XML representations. However, while textual content may be well represented in XML, much of an original document's formatting and layout might be lost as a result of the conversion process. XML is also very limited in its support for non-textual data such as photographs and graphics.

OASIS Open Document Format for Office Applications: this is an open, XML-based, format for office files, such as word-processed documents or spreadsheets. It has been adopted as an international standard (ISO/IEC 26300) and offers a suitable format for the preservation of digital documents created in proprietary office formats like those generated by Microsoft Office.

Portable Document Format Archive (PDF/A): this is a constrained version of Adobe's PDF version 1.4 which has been adopted as an international standard (ISO 19005-1). It is preservation-friendly in that: its specification is openly available; it eliminates elements likely to complicate decoding and accelerate obsolescence (e.g. audio and video elements, or encryption, etc., which are sometimes used in other PDF formats); it is self-contained (i.e. can be displayed without any reliance on information from external sources); and support for embedding metadata is very good. Records saved to this format have a look and feel which is fundamentally one of text and images designed to fit a particular page size. However, it preserves static visual appearance only, so it is not suited in cases where functionality or logical structure needs to be preserved.

Examples for images include:

Tagged Image File Format (TIFF): a format used for raster (i.e. pixel-based) images. It is widely adopted and supported by most image processing and viewing applications, and it supports sophisticated colour management features. Many repositories consider TIFF to be the best option for preserving images, and it is often used to store archival masters of digitised images. There are various sub-types of TIFF. Uncompressed Baseline TIFF (Revision 6) should be used as other revisions have additional functionalities which hinder preservation.

Joint Photographic Experts Group (JPEG): a widely used format to represent continuous tone images (e.g. photographs and greyscale images). It is defined by an international standard (ISO 10918). As with TIFF, there are different JPEG profiles; the lossless version of JPEG is preferred for preservation purposes, and the JPEG 2000, Part 1, Core coding version with lossless compression is also a favoured option.

There are also widely accepted options for sound formats (e.g. WAVE LPCM or MP3_FF) and moving image formats (e.g. MPEG-2, MPEG-4_AVC). Many digital repositories publish details of the preser-

vation formats they support, where more information about accepted formats can be found.

An important consideration when selecting a preservation format is how successfully the chosen format embodies the essential attributes, or significant properties, of the original digital object.

Significant properties

Significant properties are those aspects of the digital object which must be preserved in order to ensure that it remains accessible and meaningful over time as it is moved to new technologies. Crucially in an archival context, preserving a digital object's significant properties can also help safeguard its continuing authenticity and integrity.

Digital preservation research projects and testbeds (most recently the Investigating the Significant Properties of Electronic Content Over Time (InSPECT) Project) have categorised the significant properties of digital objects into five areas:

- Content.
- Context (metadata).
- Appearance (e.g. layout, colour).
- Behaviour (e.g. interaction, functionality).
- Structure (e.g. pagination, sections).

The inclusion of content and context here is important. While it is possible to identify some significant properties which apply to a particular format as a whole, e.g. binary word processed documents, this does not take into account the wide range of purposes for which a word processed document might have been produced – in other words the record type or genre; for instance, records as diverse as committee minutes or reports, an author's manuscript draft of a literary text, the master copy of a letter, or a list of contact details might all be saved in word-processed format. Other significant properties will therefore be object- or document-specific.

In some cases it may be decided that the textual content is the most important element of a record, in which case properties like font type and size, italicisation, pagination, layout and so on may not be essential to the document's meaning. In other cases it may be that the record creator has made use of font size, layout, bulleting, italicisation, or colour to convey or emphasise meaning, in which case these elements should be preserved. Factors which determine what elements of a record are important relate to the intention of the record creator and the requirements of the research community it is being preserved for.

The identity of the record creator is also an important factor when determining significant properties. For instance, a politician might draft some notes for a speech using a word processor; in this case it is likely to be the content of the text itself which is considered most important for preservation because the document was created as a reference tool to use whilst speaking. Alternatively, a writer might draft a poem using a word-processor; in this case, the precise layout of the text on the page will be essential to the meaning of the text and any reformatting may destroy this meaning.

Similarly, a low-res digital snapshot intended as an informal record of a holiday will gain nothing by saving it at a higher resolution, while a digital image taken by a professional digital photographer should be saved at a high resolution to maintain the quality of the photograph. Taking this approach, the Library of Congress have developed five categories of still images which are likely to be added to the Library's collections; these range from pictorial expressions of high value, such as works by graphic artists, photographers and advertisers, for whom the designated community (the users of the resource) has high interest in the artist's intent, to images incidental to web harvesting. For each category proposed, a list of preferred and acceptable formats is compiled.

Although it may not be practical or cost-effective to identify significant properties on an object by object basis, in some cases this will be necessary, especially for collecting institutions which take in the personal digital archives of a wide range of record creators. It is hoped that projects like InSPECT will be able to establish certain significant properties which are common to a particular format or type of digital object, and archivists may then be able to draw out some generalisations about particular types of record creator (e.g. politician, writer, scientist). However, there will always be specific and local considerations, and decisions will always involve an element of subjectivity. The advantage of working with donors and depositors at an earlier stage in the records lifecycle is that decisions can often be made in consultation with the record creator (see Chapter 03 *Working with record creators*).

Representation Information

Whilst information about a digital object's file format is essential for its preservation, more than this is needed to ensure that the bitstream can be transformed into something which is meaningful and understandable over time. Elements like operating system and hardware dependencies, character encoding, algorithms, standards and so on should also be taken into account. The OAIS Model uses the term Representation Information to define this kind of information. Representation Information is subdivided into three classes:

- Structure Information: describes the format and data structure concepts to be applied to the bitstream, which result in more meaningful values like characters or number of pixels.
- Semantic Information: this is needed on top of the structure information. If the digital object is interpreted by the structure information as a sequence of text characters, the semantic information should include details of which language is being expressed.
- Other Representation Information: includes information about relevant software, hardware and storage media, encryption or compression algorithms, and printed documentation.

In an OAIS information package, the Content Information (i.e. information about the digital object which is the target of preservation) is comprised of:

- The Content Data Object itself (i.e. the bits of which the object is comprised).
- The necessary Representation Information to make the content understandable to the Designated Community (the body of users who may need to access and use the digital resource). As with file formats, the Representation Information for a digital object should allow the recreation of all the significant properties of the original digital object.

A digital repository should retain persistent Representation Information along with the data objects it preserves, or it should refer to Representation Information held externally in a reliable repository.

Representation Information may need to be interpreted using further Representation Information in order to make it intelligible, e.g. it may be stated that the digital object to be preserved conforms to the ASCII standard; this standard in turn then needs to be explained. The recursive nature of Representation Information results in a complex and extensive network of representation objects, which continues expanding until the contents of the original digital object are displayed in a form the user can understand. The user in this case is a member of the repository's Designated Community (or primary user base). If this user base is small and specialised, only a minimum amount of Representation Information may be necessary. However, a repository must consider future developments and decide whether or not to maintain a larger amount of Representation Information which would render its holdings understandable to a wider community with a less specialised knowledge base. The latter is the more appropriate approach for a collecting institution which takes in personal archives; this means an extensive quantity of Representation Information is likely to be necessary.

The Digital Curation Centre has recognised that a collaborative model for creating, storing, maintaining, accessing and using Representation Information is necessary to assist the development of long-term digital curation strategies. The Centre is therefore developing a distributed Representation Information Registry/Repository[1] to provide an infrastructure for the preservation of Representation Information. The DCC will not fully populate the registry itself, so the community will only derive benefit from the registry if its members invest time and effort in populating the resource. It is intended that the registry will include:

- A structure repository containing information about file formats (with an emphasis on formats used for automated processing rather than more common formats which are adequately documented elsewhere).
- A semantic repository containing relevant data dictionaries and ontologies.

Other Representation Information to support both migration and emulation preservation strategies will also be held, such as details of software with appropriate emulation capabilities. Digital repositories will be able to refer to Representation Information held in the registry by means of a Representation Information label (in the form of an XML Schema) which can be attached to a digital object.

Selecting the right preservation strategy

There are various theories on the best way to preserve digital material, and a number of different approaches have been developed, which in turn have variants. They range from preserving the original technology on which the archival digital objects ran, to preserving only the significant properties of an object, which are defined independently of any specific hardware or software platform. Each approach has advantages and disadvantages.

Thibodeau[2] suggests that a digital archive should take the following four criteria into consideration when selecting a preservation strategy:

- Feasibility: possession of hardware and software capable of implementing the chosen method.
- Sustainability: the method should be capable of being applied indefinitely into the future; or there should be another path which will offer a sequel to the method if it ceases being sustainable.
- Practicality: implementation should be within reasonable limits of difficulty and expense.
- Appropriateness: the chosen approach should be appropriate for the particular types of digital objects to be preserved and the objectives of their preservation.

In particular he emphasises appropriateness and applicability as important factors in deciding on an approach:

- Decisions about appropriateness should be based on an informed evaluation of the significant properties of the object(s) to be preserved.
- Some preservation methods only apply to specific hardware or software platforms, some to individual data types or formats, while others are very general. Depending on the range and variety of digital objects to be preserved by a repository, selection of approach might be limited to methods that are optimal for this range, or (if very wide ranging) a method with broad applicability should be chosen.

1 Digital Curation Centre, *Representation Information Registry Repository website.* URL: <http://registry.dcc.ac.uk/omar/>

2 Kenneth Thibodeau, 'Overview of technological approaches to digital preservation and challenges in coming years', The State of Digital Preservation: An International Perspective, Conference proceedings (July 2002). URL: <http://www.clir.org/PUBS/reports/pub107/pub107.pdf>

The two principal competing strategies are migration and emulation. These two approaches and some of their variants are considered here.

Migration

Migration is the preservation approach which has been most widely practised to date. At its simplest it is defined as the copying or conversion of digital objects from one technology to another, whilst preserving their significant properties. Migration focuses on the digital object itself, rather than its environment; it aims to change the object in such a way that hardware and software developments will not affect its accessibility. It therefore applies to:

- Hardware: copying digital objects from one generation or configuration of hardware to another.
- Software: transferring digital objects from one software application or file format to another.

Whilst the OAIS Model defines refreshing as a form of migration, refreshing is essentially a means of mitigating media degradation and obsolescence, whereas full migration is also intended to overcome obsolescence of the encoding and format of the data as well. Migration (as it is generally known) is called 'Transformation' by OAIS, which defines it as changing the Content Information (the Content Data Object and its Representation Information) while attempting to preserve the full information content. When an Archival Information Package (AIP) is migrated in this way, a new 'version' of the AIP is said to have been created as a replacement for the original; both AIPs can be preserved, but the migrated version will be considered the primary package for preservation, although future developments may still result in a new migration from the original AIP. Within OAIS, the Preservation Planning function is responsible for developing migration plans, whilst performing the migration is the responsibility of the Administration function (see p. 3).

It is important that migration is fully documented by metadata, and ideally it should also be reversible: the only way to guarantee that no information will be lost on migration is to carry out a backwards migration, which should result in an exact recreation of the original object. In reality, however, some degree of 'acceptable loss' may be an inevitable result of migration; digital curators must strike a balance between achieving a perfect reversible migration and maintaining an accessible version of the digital object which is as close to the original as possible in all essential respects, but which may have undergone some subtle changes during migration. A repository might choose to define what constitutes acceptable loss for each object type as part of its content model. These are two reasons why the preservation of bitstreams is so important: if migration is unsuccessful, the repository will always have backup copies of original bitstreams to fall back on; and if some degree of acceptable loss has occurred to a digital object during migration, the migrated version can be maintained as the principal access version, while preserving its bitstream allows for the possibility of undertaking a more successful migration in the future if preservation techniques have advanced.

Migration is a very diverse field and many variations on the general approach have been considered in the digital preservation community. Four of these are discussed below:

Backwards compatibility

Simple version migration is common in the world of commercial software and has been in use for years. Successive versions of particular proprietary file formats will define linear migration paths for files stored in those formats. Software vendors usually supply conversion routines that enable newer versions of their product to read documents created in older versions and then to save in the current version. Where a digital archive includes material stored in recently-created proprietary formats which are well-supported and well-documented, a repository may decide to leave the material in this format until at risk of obsolescence; at this stage, if an upgraded and backwards compatible version of the format has been released, it may be decided to migrate to this format. However, reliance on proprietary formats like this does not provide a long-term solution, and there are major drawbacks from a digital preservation perspective:

- The updating of proprietary file formats is driven by market forces, rather than by the requirements for long-term preservation of authentic digital objects. Consequently, converting older proprietary file formats to new versions may result in the loss or alteration of important significant properties, e.g. migrations of Microsoft Powerpoint presentations to newer versions of the format can have bad results.
- Backwards compatibility is usually only limited to a few generations, and new versions of software appear on the market with great rapidity, so there is still a danger that any older documents acquired by an archive will be unreadable by the latest version of the creating software.
- The archive must have an effective technology watch function in place to monitor new developments in open and proprietary software formats held in their collections.
- The emphasis is on migrating objects to more contemporary and accessible formats rather than formats which are conducive to preservation.
- The repository must continually purchase new versions of the software.

It may also be practical to consider the onward migration of open formats to newer versions.

Migrate to standard format(s) on ingest (Normalisation)

In order to control complexity and cost, an institution may decide to support only a limited number of standardised file formats, and migrate all digital objects to an appropriate supported format on ingest. All digital objects of a particular type will be converted into a single chosen file format that is thought to embody the best overall compromise among characteristics like functionality, longevity and preservability, e.g. all raster images might be converted from their original format (such as JPEG or GIF) to Uncompressed Baseline TIFF, and all word-processed documents might be converted to OpenDocumentText (ODT). The institution then undertakes to support this format (or formats) indefinitely. This approach to migration is known as normalisation. Where it is impossible to normalise a digital object (e.g. if it was created in an obscure format), the bitstream should be preserved; if and when a tool to normalise this format is developed, it too would be subject to the normalisation process. Normalisation is seen as a more cost effective option than migrating to a wider range of formats; if using a single format like XML, repeated cyclical migration into different formats (with its accompanying risk of data loss or corruption) is avoided.

An example of normalisation which supports a limited number of preferred formats is the practice of the Public Record Office of the State of Victoria in Australia, where:

- All plain text document type records are converted to Text files which conform to specific encoding requirements.
- All formatted document type (i.e. page-oriented) records are converted either to PDF/A or to PDF (with the stipulation that the former should be used in preference to the latter).
- Images such as plans are converted to TIFF.
- Continuous tone images (e.g. photographs), especially images that do not have sharp edges, are converted to JPEG.
- Video is converted to MPEG-4.

Similarly, a repository may limit this approach to just one preservation-friendly format, converting all digital objects to this format on ingest. The National Archives of Australia takes this approach: it accepts digital records in any format and converts them all into an XML-based archival format, using the normalising tool, XENA. The XML object is the preservation master, and this is then transformed into an accessible format for users.

If normalising all digital objects to one or more standard formats, it should also be borne in mind that even open standards evolve, which means that subsequent migrations may be necessary. Even if using a single format like XML, different XML schemas will be needed for each object type;

this means that if migration becomes necessary, each schema will require a different migration pathway.

Migrate to newer or standard file formats on obsolescence

An institution may leave digital objects in their original formats and rely on their technology watch facility to identify when each format is at risk of obsolescence. For example, when a new version of software which cannot read files created in earlier versions of the software is released, all affected files are migrated to a different format. A number of migration options exist, including:

- Converting to a format considered to be a reasonable successor to the original format, e.g. a higher version of the original, although there are problems with this approach if the format is proprietary (see backwards compatibility, p. 235).
- Converting from a proprietary format to an equivalent open format which is more preservation-friendly, e.g. converting a Microsoft Word document to OpenDocument format.
- Converting to a small number of preferred formats (see normalisation, p. 236).

Migration on request

An institution may choose to leave digital objects in their original formats, or preserve only the bitstream, until a user (e.g. a cataloguing archivist or reader) requests access to them, at which point they will be migrated to a preferred format.

This approach was advocated by the Curl Exemplars in Digital Archives (Cedars)[1] and the Creative Archiving at Michigan and Leeds: Emulating the Old on the New (CAMiLEON)[2] projects. It involves preserving the bitstream of a digital object (Cedars uses the term 'byte stream', a byte consisting of eight bits) and developing a Migration on Request Tool which is able to reproduce the intellectual content of the digital object in a different format. The tool should be developed when the format is still in a usable form and its performance should be compared against a rendering of the original object; any future modifications which become necessary as technology evolves would require a similar validation process. This means that original software, screenshots and written documentation about the original environment may also need to be preserved alongside the bitstream and the Migration on Request Tool. Migration only occurs when a user requests access to a digital object, rather than taking place on a regular cyclical basis.

This approach has a number of advantages:

- Repeated format conversions are avoided, making the approach more cost-effective.
- Migration is carried out on the original bitstream, which mitigates the risk of data loss associated with repeated migrations.
- Deferring migration in this way means that by the time an object is requested, emulation techniques may have advanced and the end result seen by the user is likely to be better.
- If a record is never requested, it is never migrated, thus saving resources.

Deferring migration may have some disadvantages too:

- Migration may be required in order to assess the historical value of the archive material and make decisions about its retention.
- Migration on request tools must be kept up-to-date for each format in the archive or the tools developed may themselves be obsolete by the time migration-on-demand is requested. This obsolescence could be of two types:
    - The target format could be obsolete and the migrated object may therefore require further migration to a preferred access format.

1 Cedars Project, *Cedars Project website.* URL: <http://www.leeds.ac.uk/cedars/>
2 CAMiLEON Project, *CAMiLEON Project website.* URL: <http://www.si.umich.edu/CAMILEON/>

    - The migration tool may itself be obsolete, requiring an emulator.
- It requires preservation of software, so there may be some licensing issues.
- Cataloguers and readers may have to factor in the time needed to undertake migration into their working timetables.

Some general advantages and disadvantages of migration

Advantages

- It is a widely used strategy and procedures for simple migration are well established.
- It is generally a reliable way to preserve the intellectual content of digital objects and is particularly suited to page-based documents.
- Conversion software for some formats is readily available.

Disadvantages

- It requires a large commitment of resources, both initially and over time. Migration at the point of obsolescence is labour intensive unless it can be automated, because formats evolve so rapidly; as collections grow, the work involved in migration also increases. The migration on request approach may mitigate this to some extent, in that migration is not carried out on digital objects which may not be used; standardisation of formats also makes batch migration easier.
- Some of the data or attributes (e.g. formatting) of the digital object may be lost during migration; the authenticity of the record may then be compromised. In particular, there is likely to be a significant loss of functionality in the case of complex digital objects. Migration is based on the assumption that content is more important than functionality or look and feel.
- The potential loss of data and attributes may compromise the integrity and authenticity of a digital object, which is a major issue for digital archivists.
- There may be potential IPR problems if either the source or the new format is proprietary, although these are unlikely to be as prohibitive as they might be in the case of emulation. It is unclear yet whether the *Gowers Review*, published in December 2006, will mitigate the problem of IPR: Recommendation 10b of this report states that by 2008 libraries in the UK should be enabled to format shift archival copies to ensure that records do not become obsolete.
- Specialised conversion tools are needed to convert digital objects from one format to another, and if no appropriate tool is available for a specific file format, developing a customised migration system can be complex and expensive, although costs could be shared with institutions wishing to perform the same migration.

Emulation

In contrast to migration, which focuses on the digital object itself, emulation focuses on the technological environment in which the object was created. It involves preserving the bitstream of the object and creating an access version by using current technology to mimic some or all of the environment in which the original was rendered. This involves emulating any of the following:

- Applications: writing a new software application to do what an earlier application did; this allows files to be read on an operating system other than the one for which the original creating application was written.
- Operating systems: enabling all the software which ran on that platform to run on the emulated version. However, emulating an operating system also requires having, or emulating, an appropriate hardware platform.
- Hardware architecture: emulating hardware means that all the operating systems and applications that ran on the original hardware platform can be run without modification on the new, emulated platform. This is the most commonly employed approach; it means that the emulation of applications and operating systems is unnecessary, because they can be preserved in their original state and run on the emulated hardware.

As with migration, there are various different approaches to emulation, although many of these have not yet been widely tested in a preservation context. However, emulation itself is not new; it is a well-accepted technique that has a history of use in computer science; for instance, emulators are often developed by manufacturers to try out a design before it is produced. This involves the emulation of one computer that runs on another computer manufactured by the same vendor, but a number of emulators have also been produced to emulate one type of computer on another, e.g. emulators which run the Macintosh Operating System under Microsoft Windows on Intel-based machines and vice versa. Emulation is also widely used in computer games, and many emulators can be found on the Internet.

Emulation has great potential for long-term digital preservation. Where significant properties include things like functionality or look and feel, emulation may be a better approach than migration, which cannot guarantee to preserve functionality through changes of format. This will be most relevant to complex digital objects like websites, but in some cases the user may wish to experience the look and feel of the environment in which relatively simple objects were created, such as literary works written in a word processor. To give a truly authentic experience of the creating environment, emulation must extend to specifics like execution speed, display resolution, colour and any input devices like a keyboard or mouse, although it is debatable whether some users would want to limit the usability of an archive in this way.

Working with emulators might involve a considerable learning curve for the researcher, who may have to master the conventions of a completely unfamiliar technological environment in order to experience the object in its 'original' form. The Dutch Digital Preservation Testbed Project[1] argued that this has its equivalent in traditional archives, where a researcher may have to travel to an archive and learn to decipher unfamiliar scripts and dialects in order to work from the 'original' authentic document. It also suggested that access for less specialist researchers could be made easier by the provision of comprehensive 'help' systems, or by creating 'vernacular renditions' of the original; these are the equivalent of traditional surrogates like photocopies or transcripts and would enable the user to view the original record in a modern, understandable form.

The same project advocates the first of the three emulation approaches considered here.

Software emulating hardware

This approach is aimed at enabling the technology of the future to emulate the original computer on which the creating software ran. In order to achieve this, three elements are preserved:

- The original file.
- The software suite that rendered it.
- The hardware that ran the software.

The first two of these are preserved in the form of bitstreams. An emulator program is written to preserve the third as another bitstream; this should be written while the original computer is still extant, so that the emulated hardware platform can be validated against the computer it emulates.

There are various options to ensure that future computers can run the emulator program, e.g. 'chaining': an emulator of any one computer is able to run indefinitely once it has been implemented on one other, successor, computer; it may subsequently become necessary to emulate the successor computer in order to continue access to the original emulated hardware, resulting in a chain of emulators.

Virtual machine approach

Another variation of emulation which has been widely discussed in the digital preservation community is based on the concept of a virtual machine. Whereas a traditional emulator mimics an earlier machine which actually existed, a virtual machine emulates a computer that has never actually existed as hardware. Programs are written to run on the virtual machine rather than on a

1 Digital Preservation Testbed White Paper, *Emulation: Context and Current Status* (The Hague, June 2003). URL: <http://www.digitaleduurzaamheid.nl/bibliotheek/docs/White_paper_emulation_UK.pdf>

specific computer; if the virtual machine is then implemented on many different computers, all of the programs written for the virtual machine can run on any of those computers.

The Java Platform is a widely used example of a virtual machine: Java programs can be written to run on the Java Platform, which can be hosted on many different real computers and run identically on all of them. While the Java virtual machine is unsuitable for long-term preservation (it evolves rapidly, meaning that it is relatively unstable, and its language is specific to Java), it would be possible to produce an 'Emulation Virtual Machine' for digital preservation purposes. If such a machine could be defined, it would then become the virtual platform on which all emulators are written to run. As each computer in a given generation reaches obsolescence, an emulator of that computer is written to run on the virtual machine.

Universal Virtual Computer (UVC)

This is a variant on the virtual machine approach, which has been developed by R.A. Lorie of the IBM Research Centre in Almaden, USA in 2004, in conjunction with the Koninklijke Bibliotheek in the Netherlands. It involves preserving the bitstream of a digital object along with a specially written emulation program. This program is designed to run on future computers and to emulate the computer on which the digital object was created; it is written in the simple machine language of a platform-independent UVC, and a future computer would need a UVC Interpreter to read and execute the program. Once this is done, the original bitstream could be accessed – by means of the UVC – on any future computer.

In theory, UVC programs can be written for each file format. The UVC Interpreter deciphers each program into a Logical Data View, possibly an XML-like structure, which describes in detail how the digital object is structured, e.g. raster-based images are described pixel by pixel. This is then translated into an understandable representation for the user. This latter stage obviously involves an element of migration, so the UVC essentially combines both emulation and migration approaches.

The UVC method has been successfully applied to images stored in GIF, and the Koninklijke Bibliotheek hopes to extend this to TIFF and PDF. The UVC must support a much wider range of formats if it is to become a central plank of a preservation strategy.

Some general advantages and disadvantages of emulation

Advantages

- In theory full emulation enables us to recreate the full functionality and exact look and feel of a digital object's performance. It is therefore an attractive approach for preserving complex digital objects and those where appearance or functionality are identified as significant properties.
- In contrast to migration, the focus of emulation is on changing the environment rather than the digital object itself, thus lessening the risk of data loss through repeated migration cycles.
- Oltmans and Kol have concluded that emulation is more cost-effective for preserving large collections, despite the relatively high initial costs for developing an emulation device; in contrast, migration applies to all the objects in a collection repetitively, creating high ongoing costs. However, the need for chaining emulators in the future may detract from this.[1]
- The emulation approach can be implemented at a higher level than the migration approach, so rather than developing conversion solutions per format institutions can develop emulation solutions per environment.
- It means that records in obscure formats do not have to be abandoned; in theory if the creating hardware/software can be emulated, all the records created in that environment can be recreated.
- Regardless of the principal preservation approach adopted by a digital repository, emulation could be useful as a backup mechanism that would provide access to the 'digital original' form of each record and may be necessary for the extraction of digital objects from older technological environments.

1 Erik Oltmans and Nanda Kol, 'A comparison between migration and emulation in terms of costs', *RLG DigiNews*, 9, 2 (15 April 2005). URL: <http://www.rlg.org/en/page.php?Page_ID=20571&Printable=:1&Article_ID=1714>

Disadvantages

- As yet, emulation has not been widely tested as a long-term digital preservation strategy, and further practical tests are essential before more definitive conclusions about its reliability can be drawn.
- An emulation system may require the user to master completely unfamiliar technology in order to understand an archival digital record, and technological developments are incredibly rapid; for instance, many have already forgotten how to use relatively recent word processing programs like Wordstar. This problem could potentially be addressed by developing different means or levels of access.
- Selecting an emulation strategy also involves buying into a migration strategy because emulators themselves become obsolete, so it becomes necessary to replace the old emulator with a new one, or to create a new emulator that allows the old emulator to work on new platforms.
- Most emulation approaches will involve preserving or emulating proprietary software which is covered by patent, licence or other IPR. This is a major issue and must be addressed by any institution introducing an emulation strategy; it is unclear yet whether the *Gowers Review* (see p. 263) will alter this situation.
- The concept of 'exact original look and feel' is itself debatable; can it therefore be preserved by emulation? Digital objects are so dependent on the environment used to render them; for instance, a user's experience of a website can differ according to what software and hardware they are using.
- Emulation may require a large commitment in resources, and highly skilled computer programmers would be needed to write the emulator code.
- If the UVC approach is used, large numbers of decoder programs will be necessary to cope with the variety of file formats that are available, and it may be that new UVC emulators need to be written for each new generation of hardware.

Other preservation approaches

Encapsulation

Encapsulation is an essential element of many emulation approaches and also plays a key part in some other preservation strategies. It involves retaining a digital object in its original form as a bitstream, and encapsulating it along with instructions and whatever else might be necessary to maintain access to it in the future; this might include software viewers or software specifications for emulation, as well as comprehensive preservation metadata.

An Information Package as defined by the OAIS Model represents a form of encapsulation, in which the digital object is packaged together with: the Representation Information needed to interpret the bits appropriately for access; and Preservation Description Information, which includes information on provenance, context, reference and fixity.

The Virtual Machine approach is an extension to encapsulation in that an executable program is also packaged together with the digital object.

Technology preservation

Like emulation, this approach focuses on the technological environment rather than on the digital object. Instead of mimicking the original environment, it involves preserving the digital object together with all the actual hardware and software required to maintain access to the object; this includes operating systems, original application software and media drives. It could be argued that maintaining the original technology is the most effective and obvious means of preserving the look and feel of a digital environment, and there is certainly merit in keeping samples of old computer systems as a resource for researchers in the future; however, while it might offer a short-term solution, this is not a viable strategy for long-term digital preservation, for various reasons:

- Cost and space implications for acquiring and maintaining large quantities of hardware (from computers and peripherals to connectors) are prohibitive for many organisations.

- Older operating system and application software and appropriate licences must also be acquired and maintained.
- Over time the machines will degrade and ultimately fail, so the number of machines capable of reading certain types of old files will continually decrease.
- Technical support for both software and hardware will also disappear over time.
- Documentation for older computing environments can be difficult to locate.

Digital archaeology

Digital archaeology involves retrieving data from obsolete software or hardware environments, and obsolete or damaged media, such as punch cards, 8" floppy disks and the wealth of other removable media which have been used since the earliest days of computing. There are a growing number of specialist third party services offering to carry out digital archaeology, and it has been shown to be technically possible to recover bitstreams from damaged and obsolete media. Only trained specialists will be able to extract data in this way, using special hardware and software; for instance, in order to extract data from relatively recent, damaged, media, the British Library makes use of 'forensic' hardware, designed for use by law enforcement, intelligence, corporate and military agents who need to recover digital evidence from hardware in a way which ensures its authenticity.

Digital archaeology will form an inevitable part of the digital archivist's work for a long time to come; however pro-active our approach in working with donors and depositors, archives are still likely to contain obsolete or damaged data which we need to rescue from oblivion. Ultimately, however, digital archaeology is an emergency recovery strategy, not a pro-active and preventative approach to long-term preservation, because:

- It is much more costly than the other major preservation strategies and is unlikely to be cost-effective for any other than the most highly valued digital resources.
- Relying on digital archaeology means that the digital material which isn't necessarily highly valued (yet might still be useful to some researchers or have important evidential value) might not be rescued.
- If there is no accompanying metadata or documentation, it may be impossible to assess the value or usefulness of obsolete digital resources until after rescue has taken place, which may turn out to be a waste of resources.
- Digital archaeology techniques are unlikely to be successful in all cases.
- It requires a certain amount of technology preservation (see above).

If digital archaeology is successfully carried out as an emergency rescue measure, digital archivists must then process extracted files using the repository's policy and procedures and decide whether or not to retain the original media on which the creator stored their files. Some arguments in favour of retaining media include:

- The value of removable media as artefacts, especially if the record creator has annotated media labels.
- They constitute the 'original' version of the digital object.
- The future researcher may wish to experience the original environment.
- Future developments in technology may mean that more data can be extracted from the media.

Arguments against retaining media:

- Disks and other removable media are equivalent to the original files or packaging in which traditional archives arrive at the repository. Unless there is special significance in these (e.g. they have particular evidential value or there is a possibility that all data has not been extracted), the usual approach is to transcribe any original labels and discard the original file cover; there is no reason to treat digital media any differently.

- The media will fail with time, as will the devices needed to read them, and maintaining the necessary hardware to read old media involves all the disadvantages associated with technology preservation.
- Successful emulation is a better way of enabling the future researcher to experience the original environment.

Preservation strategies for personal digital archives

The digital archive material acquired from working politicians and used as Paradigm's primary testbed is comparatively small in quantity, yet contains a very wide range of file formats. During the first 10 months of the project alone, the material accessioned included 20 different formats, and this is only material which was created during the last five years. Forms encountered include email, word-processed documents, spreadsheets, digital images, presentations, personal web pages and blogs. Paradigm also worked with the digital component of the archive of Barbara Castle, which represented three generations of computing technology – hardware, software and formats.

This is likely to be typical of personal digital archives generally. Institutions that collect personal archives can have little control over the formats, software and hardware used by record creators. Curators of personal digital archives have recognised the need to work more closely over long periods with their donors and depositors, and to offer guidance on managing digital material – like the guidelines (see *Appendix B: Guidelines for creators of personal archives*) created by Paradigm; however, it is likely that collecting institutions will always have to deal with a wide variety of digital material, at least some of which may be obsolete. This means that some level of digital archaeology will be necessary for many years to come.

In such a diverse environment, it is necessary to select very broadly applicable preservation strategies, or to consider implementing a combination of different approaches:

- An institution's preservation strategy should define the range of formats to be supported and include detailed information about how each format will be treated; this will involve identifying significant properties and creating profiles (content models) for different object types, as well as taking into account the different categories of record creator represented in the collections.
- The enormous variety of formats to be dealt with, and the range of creators involved, may mean that normalisation is inappropriate as an overall approach because significant properties are likely to be widely varied and very specific. However, normalisation may be the most suitable way of dealing with digital objects in old or obscure formats; alternatively these might be preserved at bit level only.
- Migration to standard formats on obsolescence may be an appropriate strategy for dealing with objects in well-supported current formats. A comprehensive technology watch programme should always be in place if taking this approach.
- Migration on request might be a useful approach for personal digital archives, which will often be closed for extended periods under the *Data Protection Act* and copyright (see pages 250 and 253) restrictions; unnecessary migrations during the closure period would be avoided and reduce the risk of data loss.
- Although not widely tested, emulation offers a promising solution for preserving complex objects and those where maintaining look and feel is a priority; the latter may vary according to the type of record creator concerned.
- Repositories seeking to engage seriously in the preservation of personal digital archives should start collecting hardware, software and manuals for data extraction and digital archaeology.

- Bitstreams should always be preserved intact, and be subject to preservation measures like refreshment, backup and fixity checks. They should be packaged together with appropriate metadata, including Representation Information. This can be achieved by creating a METS document for each digital object which contains or refers to representation information and other types of preservation metadata. Representation information is likely to be extensive for collecting institutions which are used by a wide range of researchers.
- Curators should test available migration and emulation tools on digital material they already hold in order to develop informed preferences and to assist in the development of such tools.
- All preservation actions should be fully documented. Paradigm explored and recommends PREMIS (also stored in a METS document) as a means to record information about any migrations or other preservation activities.

Whatever combination of approaches is selected, the result must be affordable. The strategy developed must embody the best that can be accomplished with the available resources. Collaboration and knowledge-sharing will be vital to establishing a successful long-term preservation strategy for personal digital archives; evolving strategies in-line with the digital preservation community will allow repositories to leverage tools and techniques developed by others. The first wave of digital archivists might also wish to exercise some caution by adopting multiple parallel strategies while the field is still at an early stage of development and tools are quite limited. Information technology continues to evolve, as does the digital preservation community: new projects and testbeds are being created on a regular basis, and shared registries and tools are constantly developing and growing. This environment offers the digital archivist many challenges but also a huge information resource on which to draw.

Useful resources

Arms, Caroline R. and Fleischhauer, Carl, 'Digital formats: factors for sustainability, functionality, and quality', paper given at the IS&T Archiving 2005 Conference, Washington DC (29 April 2005).
URL: <http://memory.loc.gov/ammem/techdocs/digform/Formats_IST05_paper.pdf>

Au Yeung, Tim, 'Media choices for the preservation of digital documents', *AIC News* (March 2005).
URL: <http://aic.stanford.edu/sg/emg/library/pdf/yeung/yeung-2005-07-22_au-media-choices.pdf>

Ball, Alex, *Briefing Paper: File Format and XML Schema Registries* (2006).
URL: <http://www.ukoln.ac.uk/projects/grand-challenge/papers/registryBriefing.pdf>

Brown, Adrian, *Digital Preservation Guidance Note 1: Selecting File Formats for Long-Term Preservation* (The National Archives, June 2003).
URL: <http://www.nationalarchives.gov.uk/documents/selecting_file_formats.pdf>

Brown, Adrian, *Digital Preservation Guidance Note 2: Selecting Storage Media for Long-Term Preservation* (The National Archives, June 2003).
URL: <http://www.nationalarchives.gov.uk/documents/selecting_storage_media.pdf>

Cedars Digital Preservation Project, *Cedars Guide to: Digital Preservation Strategies* (2 April 2002).
URL: <http://www.leeds.ac.uk/cedars/guideto/dpstrategies/dpstrategies.html>

Cornell University tutorial, *Digital Preservation Management: Implementing Short-term Strategies for Long-term Problems*. See chapter on 'Obsolescence'.
URL: <http://www.library.cornell.edu/iris/tutorial/dpm/oldmedia/index.html>

Digital Preservation Testbed White Paper, *Migration: Context and Current Status* (The Hague: ICTU, 5 December 2001).
URL: <http://www.digitaleduurzaamheid.nl/bibliotheek/docs/Migration.pdf>

Digital Preservation Testbed White Paper, *Emulation: Context and Current Status* (The Hague: ICTU, June 2003).
URL: <http://www.digitaleduurzaamheid.nl/bibliotheek/docs/White_paper_emulation_UK.pdf>
Enderle, Rob, 'Optical HD battle may be over', *Digital Trends* (6 December 2006).
URL: <http://news.digitaltrends.com/talkback158.html>

Granger, Stewart, 'Emulation as a digital preservation strategy', *D-Lib Magazine*, 6, 10 (October 2000).
URL: <http://www.dlib.org/dlib/october00/granger/10granger.html>

Jones, Maggie and Beagrie, Neil, 'Preservation Management of Digital Materials: A Handbook', *Digital Preservation Coalition website*.
URL: <http://www.dpconline.org/graphics/handbook/>

Mellor, Phil, Wheatley, Paul and Sergeant, Derek, 'Migration on request, a practical technique for preservation', from ECDL 2002: European Conference on digital Libraries, LNCS 2458, pp. 374-389 (Rome, September 2002).
URL: <http://www.springerlink.com/content/752vmvw0g0w40dj2/fulltext.pdf>

Lorie, Raymond, 'A project on preservation of digital data', *RLG DigiNews*, 5, 3 (15 June 2001).
URL: <http://www.rlg.org/preserv/diginews/diginews5-3.html>

Lorie, Raymond, *The UVC: A Method for Preserving Digital Documents – Proof of Concept,* IBM / KB Long-Term Preservation Study Report Series No. 4 (Amsterdam: IBM Netherlands, December 2002).
URL: <http://www.kb.nl/hrd/dd/dd_onderzoek/reports/4-uvc.pdf>

Oltmans, Erik and Kol, Nanda, 'A comparison between migration and emulation in terms of costs', *RLG DigiNews*, 9, 2 (15 April 2005).
URL: <http://www.rlg.org/en/page.php?Page_ID=20571&Printable=:1&Article_ID=1714>

PREMIS Working Group, *Data Dictionary for Preservation Metadata* (May 2005).
URL: <http://www.oclc.org/research/projects/pmwg/premis-final.pdf>

Stanescu, Andreas, 'Assessing the durability of formats in a digital preservation environment: the INFORM methodology', *D-Lib Magazine*, 10, 11 (November 2004).
URL: <http://www.dlib.org/dlib/november04/stanescu/11stanescu.html>

Stevenson, Jane, *JORUM Preservation Watch Report* (July 2005).
URL: <http://www.jorum.ac.uk/docs/pdf/Digital_Preservation_Report.pdf>

Thibodeau, Kenneth, 'Overview of technological approaches to digital preservation and challenges in coming years', The State of Digital Preservation: An International Perspective, Conference proceedings (Washington DC: Council on Library and Information Resources, July 2002), pp 4-31.
URL: <http://www.clir.org/PUBS/reports/pub107/pub107.pdf>

Wheatley, Paul, 'Migration – a CAMiLEON discussion paper', *Ariadne*, 29 (September 2001).
URL: <http://www.ariadne.ac.uk/issue29/camileon/>

Examples of institutional preservation policies:

Arts and Humanities Data Service (AHDS), 'AHDS Repository Policies and Procedures', *AHDS website.*
URL: <http://www.ahds.ac.uk/preservation/ahds-preservation-documents.htm>
Includes links to AHDS preservation policy, and preservation handbooks for different data types.

Florida Center For Library Automation, 'Florida Digital Archive', *Florida Center For Library Automation website*.
URL: <http://www.fcla.edu/digitalArchive/>
Includes policy guide, recommended file formats, and action plans for specific formats.

National Archives of Australia, 'Information management framework', *National Archives of Australia website.*
URL: <http://www.naa.gov.au/records-management/publications/topic.aspx>
Includes various documents relating to the National Archives' preservation strategy, which focuses on normalisation.

OCLC Online Computer Library Center, *OCLC Digital Archive Preservation Policy and Supporting Documentation* (8 August 2006).
URL: <http://www.oclc.org/support/documentation/digitalarchive/preservationpolicy.pdf>