4. TECHNOLOGICAL DECISIONS

4.1 What is involved in preserving digital resources?

Preservation is concerned with ensuring the longevity of a digital resource through changing technological regimes with a minimum loss of the resource's intellectual content. Since the format and functionality of a resource in a particular computer environment contribute to its intellectual content, there is more to preservation than simply 'preserving the information', whether that information is expressed as a set of images, a textual document, or a multimedia document.

Three main approaches to digital preservation have been developed:

- Preserve the original software (and possibly hardware) that was used to create and access the information. This is known as the *technology preservation strategy*. It also involves preserving both the original operating system and hardware on which to run it.
- Program future powerful computer systems to emulate older, obsolete computer platforms and operating systems as required. This is the *technology emulation strategy*.
- Ensure that the digital information is re-encoded in new formats before the old format becomes obsolete. This is the *digital information migration strategy*.

These three approaches are described in more detail below.

Those involved in digital preservation must handle a wide range of digital resources and these will influence their choice of digital preservation strategy. Some strategies are more suited than others to preserving a particular resource — the choice of strategy must reflect fitness for purpose. Table 6 lists ten generic digital resource categories and their characteristics. It is not a definitive grouping, but a selection of the common digital resource categories which those involved in the creation of and access to digital material are likely to encounter.

Table 6. Categories of digital resource

	Category of digital resource	**Data types included in resource**	**Applications used to create/manage/distribute digital resource**	**Notes**
1	**Data Sets**	Alphanumeric data	Wide range of data processing applications; bespoke software and application packages; managed in flat file; networked; hierarchical; relational and object oriented databases; presented via presentation graphics, modelling software, report writers, etc.	Survey data; results of experiments; transaction data; event data; administrative data; attribute data; bibliographic data

2	**Structured Texts**	Alphanumeric data; mark-up data; tags to other data types (raster and vector graphics)	Word processing; text editing; HTML editors; desktop/corporate publishing; LaTeX; SGML and application specific Document Type Definitions; XML	Literary texts; formal documents; corporate publications; commercial publications; Web pages
3	**Office Documents**	Alphanumeric data; mark-up data; raster and vector graphics	Word processing; spreadsheets; document image processing; office suites; groupware; document management systems; relational databases	Sets of digital documents; Digitised paper images; links/bundles created via office suites; groupware; HTML
4	**Design Data**	Vector and raster graphics; alphanumeric data	CAD; word processing; document image processing; relational databases; object oriented databases	Product data; as built drawings; Models; plans
5	**Presentation Graphics**	Vector/raster graphics; moving graphics alphanumeric data; full motion video; interleaved audio and video	Business graphics; clip art; creative graphics; present-ation systems; Computer Based Training; multimedia	Business presentations; formal courseware; CBT packages
6	**Visual Images**	Raster graphics; alphanumeric data	Image capture software; image processing and editing software; object oriented; relational and flat file databases	Fine art; picture libraries; photographic libraries; medical images; images of historic/manuscript documents
7	**Speech & Sound Recordings**	Audio data; MIDI; metadata	Speech processing; audio recording and playback; symbolic music recording; relational and flat file databases	Music libraries; sound effects; radio broadcasts; sound recordings; media
8	**Video Recordings**	Digital video; full screen, full motion video; interleaved audio and video; metadata	Digital video frames stored as bitmaps; audio files; audio/video interleaved; compression systems; Relational & Flat File Databases	Media libraries; training centres; video clips games
9	**Geographic/ Mapping Data**	Vector and raster graphics; alphanumeric data	GIS systems; mapping software; relational & object oriented databases	Maps; coordinates; range of overlay data; links between data types
10	**Interactive Multimedia Publications**	Interleaved audio and video data; moving graphics; vector and raster graphics; alpha-numeric data	Authoring software; editing software; access software	Electronic publishing; educational and training material; marketing material; games etc

Certain technical factors will also impact on the choice of strategy for each category of resource: the basic data types employed in each category; the application programs used to create them; the structures applied to them; and the systems used to manage or distribute them prior to deposit.

4.2 Technology preservation

Being able to replicate the behaviour of a program and the look and feel of a document or publication has to be balanced against the costs and the technical difficulties that would be faced by anyone trying to keep ageing computer hardware platforms running.

Already in the brief history of computing, hundreds if not thousands of old proprietary computer hardware platforms have disappeared without trace. Today we are seeing the increasing dominance of a few computer platforms and, in theory, this should simplify the task of preserving them in future. However, even here there are difficulties, given the rapid obsolescence of computer components. It is unlikely that components for today's PCs could be sourced in ten years' time.

In general, this strategy can only be regarded as viable for the short to medium term, as a relatively desperate measure in cases where valuable digital resources cannot be converted to hardware and/or software independent formats and migrated forward. This would usually be due to the complexity of the digital resource and the fact that it was created on a proprietary and obsolete application program. Any collection manager in charge of a large collection of digital resources who relied solely on this strategy would very soon end up with a museum of ageing and incompatible computer hardware.

4.3 Technology emulation

This strategy involves designing emulator programs on current and future computer platforms and programming them to mimic the behaviour of old hardware platforms and to emulate specific operating system software. It requires extremely detailed specifications for the outdated hardware and operating system software. A less comprehensive approach might be to emulate specific data operations as, for example, programs set up to read obsolete world processing formats and display the information.

In general, emulation is a specialist strategy, where the need to maintain the look and feel of the original digital resource is of great importance to the users of the collection. As with technology preservation, it would be used in cases where digital resources cannot be converted into software independent formats and migrated forward.

Anyone relying solely on this strategy could be taking a significant risk. They would be depending on the technical ability of the software engineers to emulate a specific environment and sustain it, and on the commercial viability of anyone providing such a service.

4.4 Digital information migration

Behind this strategy is the assumption that it is only worth preserving digital information if you can access it on current computer hardware and software platforms.

As those platforms change, so collection managers must migrate their digital resources forwards to ensure that they remain accessible on the new platforms.

According to the Task Force on Digital Archiving (Waters and Garrett, 1996), migration is not optional — it is an essential operation, given the fact that the computer operating environments of digital archives will inevitably change over time.

There are a variety of migration strategies, appropriate to different formats of digital materials. The Task Force makes the point that no single strategy applies to all formats and none of the current preservation methods is entirely satisfactory.

A basic migration strategy involves **changing media** — transferring digital resources from less stable to more stable media. The simplest version of this strategy involves moving information from fragile magnetic media to more stable, controlled environments. A more extreme version would involve printing digital information onto paper or recording it on preservation quality microfilm. While this preserves a copy of the basic data, it destroys the digital functionality of the resource. Much valuable data — for example, computation capabilities, graphic display or indexing — will be lost in the process. It is impossible to microfilm the equations embedded in a spreadsheet, to print out an interactive full motion video, or to preserve a multimedia document as a flat file. Changing media should therefore be regarded as a last resort if no other strategy meets the requirements.

A second migration strategy relies on popular application software having **backward compatibility**. The latest versions of most popular word processing packages are capable of decoding files created on earlier versions of the same package. Migration involves testing the process, loading files into the new version and saving them in the new file format.

However, this strategy cannot be relied upon over the long term or for more complex digital resources. No one software supplier is in control of all the technical or commercial factors needed to guarantee the continued viability and support of their application software. They may go out of business, or introduce a totally new software package and drop support for their old package.

The third migration strategy relies on **interoperability** between rival popular application programs. Digital resources created on one application program can be exported in a common interchange format and then imported into a rival application program without the need to run the specific program that was used to create them.

If such interoperability could be guaranteed between all the major competing application programs, then digital information migration would be a much easier process! The list of migration options would be endless.

Today, similar software can, at least partially, interpret files created by a different software package. However, even with simple digital resources the interchange is likely to involve some loss of data. For example, word processing programs allow authors to save work as simple alphanumeric text using the American Standard Code for Information Interchange (ASCII) or other interchange formats such as Rich Text

Format (RTF). Documents are rarely pure text — they also include format data, figures and footnotes, which may be lost in the interchange.

The more complex the digital resource, the greater the potential loss is likely to be. For example, interchanging the data held in geographical information systems (GIS) databases and groupware databases could involve the loss of thousands of links that have taken years of effort to create and which represent the bulk of the value of the database.

A fourth migration strategy — **conversion to standard formats** — is particularly appropriate for digital archives with large, complex, and diverse collections of digital materials. It proposes an enhanced version of the interoperability strategy. Where the latter relies on interchange formats that can be generated automatically from within applications, with this strategy the onus is placed on collection managers to define the preferred formats and select those that are most appropriate for the digital resources they collect and the users they serve.

The aim is to migrate digital objects from the great multiplicity of formats used to a smaller, more manageable number of standard formats that can still encode the complexity of structure and form of the original. For example, a digital archive might require that textual documents conform to standards such as SGML, or that images conform to a Tagged Image File Format (TIFF) format and standard compression algorithms.

The formats chosen for conversion will be determined by the structure of the digital resources themselves, the objectives set by the collection manager, and users' requirements. For example, the decision will be influenced by whether priority is given to preserving the ability to process or edit the digital resource, or to preserving its format or visual presentation.

For example, in many data centres a key objective is to preserve and make the data available in a format which allows them to be loaded into user application programs, processed and new data derived from them. The content is the valuable resource, while its presentation is of secondary importance. In archives and records centres, on the other hand, the priority may well be to preserve the format or visual presentation of the digital resource, in order to ensure its archival integrity.

While digital information migration is widely adopted, as a strategy it is still evolving. Techniques should become more effective as practical experience grows, but it remains largely experimental. Further research and development efforts are required to test the technical feasibility of the various approaches over the long term. However, since most migration strategies have the potential to lose information, a copy of the original bit-stream should be preserved in all cases. This may later turn out to contain important clues lost or corrupted in the migration process.

4.5 Selecting the most appropriate long term preservation strategy

The best preservation strategy for **data sets** is conversion to standard formats. Some data sets are created in a 'controlled' environment, following professional standards for the production of formal interchange documents, or standard templates and house styles. If this is not the case, and the data are incomplete or lack adequate documentation, the archive may have to spend time and resources processing the data in order to make them usable.

The best approach to preserving data sets is for the archive to set standards and provide guidance for depositors, stipulating how records should be structured for deposit. The Research Councils — for example, the Natural Environment Research Council (NERC), described in the previous chapter — are beginning to adopt this approach, making it the responsibility of depositors to document their data sets fully and to deposit them in an appropriate format.

Most archives which specialise in the preservation of **structured texts** have adopted the digital information migration strategy, especially conversion to standard formats. The Oxford Text Archive (OTA), for example, follows the Text Encoding Initiative (TEI) guidelines for using SGML to mark up literary works. There are different versions of TEI for different categories of literary work; OTA uses TEI Lite as a basic standard. Increasingly, regular depositors are themselves adopting TEI Lite.

In some cases deposited texts may have been created using a proprietary desktop publishing application or management system. If these became obsolete then the archive might need to preserve the operating environment that they ran on before developing a migration strategy that allowed conversion to a standard format.

Office documents may be created in a variety of data types. The Public Record Office (PRO) draft guidelines address the long-term storage of electronic records created on office systems and recommend the digital information migration strategy, specifically conversion to standard formats. The formats they recommend are PostScript, Adobe's Portable Document Format (PDF), TIFF, SGML, and Comma Separated Variable ASCII for alphanumeric data. The PRO endorses strategies in which the format and presentation of the document are preserved as well as its content.

The PRO also identifies a potential problem with proprietary applications where, for example, vital attribute data and links between the documents might be held in an office suite application, groupware database or document management system. In this case it would be necessary to preserve the operating environment that they ran on until the links could be captured in a standard form.

Design data are made up of many data types: vector graphics (computer aided design); raster graphics (old manual drawings); and alphanumeric data (text documents plus attribute data in databases). They may be held in two and three-dimensional formats.

Archives which manage this category of digital resource recommend the digital information migration strategy and make use of a combination of backward compatibility, interoperability and conversion to standard formats. For interchange they tend to rely on the de facto DXF format, or Initial Graphics Exchange

Specification (IGES) for two and three dimensional vector graphics. Standard formatted forms include HPGL; Encapsulated PostScript and TIFF.

Presentation graphics may comprise vector graphics (CAD), raster graphics, animation (moving graphics), and alphanumeric data. The digital information migration strategy is appropriate here, using a combination of backward compatibility, conversion to standard formats and changing media. Standard formatted forms include PostScript; Adobe PDF and TIFF. Technology preservation or technology emulation would only be employed where valuable presentations had been left on a proprietary platform which was now obsolete and where valuable data would be lost in translating the data to a standard format.

Visual images comprise one main data type — raster graphics data. In addition, alphanumeric data relating to the images may be held and managed in a database. Most archives which specialise in the preservation of visual images have adopted digital information migration, specifically conversion to standard formats. Those who are involved in preserving visual images must also make decisions on resolution, file formats, compression, and whether the images should be captured as black and white, greyscale, or colour.

For the very specialised area of **speech or sound recordings** the digital information migration strategy is suitable, using a combination of backward compatibility, conversion to standard formats and changing media.

Digital speech processing can be divided up into three areas: speech coding (the analogue-to-digital conversion of speech signals or waveforms, the compression of the digital signals, and the reverse digital-to-analogue conversion for play back purposes); speech synthesis (the translation by computers of a coded description of a message into speech, i.e. computers ‘talking’); and speech recognition and understanding, which facilitate people ‘talking’ to computers, dictating text or issuing commands. Sound coding involves the analogue-to-digital conversion of sound signals, the compression and/or storage of the digital signal, and the reverse digital-to-analogue conversion for play back purposes. Any sound, including speech and music, can be recorded in this way. Music can also be described in a symbolic way — we have used printed musical scores for centuries. For computer systems the Musical Instrument Digital Interface (MIDI) standard defines how to code all the elements of musical scores, including notes, timing conditions, and the instruments to play each note. The sound files which are to be preserved will generally contain sound data coded either as a digitised analogue sound signal or as notes for a MIDI instrument.

Digital video recordings, another very specialised area, involve one main data type — motion video or moving image data. Increasingly, digital video resources contain interleaved audio data as well. Alphanumeric data relating to the moving images or interleaved audio and video data may also be held and managed in a database. The preservation of digital video data is still in the early stages and the bulk of holdings are likely to be in analogue film or videotape formats. Where practical, archives have adopted the digital information migration strategy, especially conversion to standard format. The Motion Picture Experts Group (MPEG) standards provide several standards for the compression of full motion video.

This is an area where early digital video material was created on proprietary applications and held in proprietary formats. Where the applications are now obsolete, a technology preservation or technology emulation strategy may be needed as a temporary measure to preserve the data.

Geographic/mapping data can cover almost all data types, including raster graphics, (base mapping data), vector graphics, and alphanumeric data held in databases. Geographic data can be held in two and three-dimensional formats. Cartographic software packages range in sophistication and functionality from atlases and route planners up to full geographical information systems (GIS). For this category of digital resource the digital information migration strategy is recommended, using a combination of backward compatibility, interoperability and conversion to standard formats as appropriate, with changing media as a back-up option.

By definition **multimedia** publications involve at least three data types. Most contain motion video with audio data interleaved; many comprise animation with interleaved audio data; and many will also involve some still images, graphics and alphanumeric data. Most early multimedia publications were produced on one of the compact disk (CD) formats and will have been authored and edited using proprietary multimedia editing and authoring packages. They will be accessed via proprietary access software. A strategy of technology preservation or technology emulation may be the best way to preserve the data until a practical migration strategy is developed. It may prove difficult to migrate the data in future without the loss of considerable data.

This is a very specialised and still relatively new area where individual studies need to be conducted by experts in the field who appreciate the specific challenges and risks which a migration strategy pose to interactive multimedia publications. They probably represent the most difficult category of digital resource to preserve today.

Table 7 provides a summary of digital resources and the preservation strategies recommended for them.

Table 7. Summary of digital resources and their recommended preservation strategies

	Digital Resource	**Preservation Strategy**	**Subset of Strategy**	**Notes**
1	**Data Sets**	Digital Information Migration	Conversion to Standard Formats	
2	**Structured Texts**	Digital Information Migration	Conversion to Standard Formats	
3	**Office Documents**	Digital Information Migration	Conversion to Standard Formats, Backward Compatibility, Change Media	
4	**Design Data**	Digital Information Migration	Backward Compatibility, Interoperability, Conversion to Standard Formats, Change Media	Technology preservation/emulation as short term strategy for product data on obsolete systems.
5	**Presentation**	Digital Information	Backward	

	Graphics	Migration	Compatibility, Conversion to Standard Formats, Change Media	
6	**Visual Images**	Digital Information Migration	Backward Compatibility, Conversion to Standard Formats, Change Media	
7	**Speech/Sound Recordings**	Digital Information Migration	Backward Compatibility, Conversion to Standard Formats, Change Media	A specialised area where additional work is needed by experts in the field.
8	**Video Recordings**	Digital Information Migration	Backward Compatibility, Conversion to Standard Formats, Change Media	A specialised area where additional work is needed by experts in the field. Technology preservation/emulation needed in short term where data locked in proprietary systems.
9	**Geographic/ Mapping Databases**	Digital Information Migration	Backward Compatibility, Interoperability, Conversion to Standard Formats	A specialised area where additional work is needed by experts in the field. Technology preservation/emulation needed in short term where data locked in proprietary systems.
10	**Interactive Multimedia Publications**	Technology preservation/emulation in short term for data in proprietary systems until agreed migration strategies can be developed		A specialised area where additional work is needed by experts in the field.