British Library Research and Innovation Report 106

Comparison of Methods & Costs of Digital Preservation

A Consultancy Study Conducted By
Tony Hendley, Technical Director,
Cimtech Ltd, University of Hertfordshire

British Library Research and Innovation Centre
1998

This study is part of a programme funded by JISC as a result of a workshop on the Long Term Preservation of Electronic Materials held at Warwick in November 1995.

The programme of studies is guided by the Digital Archiving Working Group, which reports to the Management Committee of the National Preservation Office.

The programme is administered by the British Library Research and Innovation Centre.

RIC/CT/316

ISBN 0 7123 9713 2

ISSN 1366-8218

British Library Research and Innovation Reports may be purchased as a photocopy or microfiche from the British Thesis Service, British Library Document Supply Centre, Boston Spa, Wetherby, West Yorkshire LS23 7BQ, UK.

1 Background & Terms of Reference for the Study

1.1 Background

In 1995 FIGIT and BLR&DD co-sponsored a workshop on the "Long Term Preservation of Electronic Materials" at Warwick University. The workshop produced a list of recommended actions which were considered by the Management Committee of the National Preservation Office and JISC subsequently agreed to fund a programme of studies to be developed by the NPO and administered by the British Library Research and Innovation Centre. The National Preservation Office's Digital Archiving Working Group drew up the programme.

1.2 Terms of Reference for the Study

This study was one of a total of seven studies commissioned as part of the initial programme.

The workshop recognised that a vast number and type of digital documents and resources have been created over the past thirty years and millions of new digital objects are being created each day. As the digital store grows and diversifies, so, too, is the preservation management challenge expanding and becoming more complex. Caretakers (librarians, archivists, information technologists) of this ever expanding body of digital cultural heritage material need effective strategies to preserve and sustain viable access to the resources already in their repositories or likely to arrive there in the future.

In the Commission On Preservation & Access/ Research Libraries Group report on digital preservation (1) and the outcomes from the November 1995 meeting at Warwick (2), three potential strategies for ensuring long-term access to digital information were identified: technology preservation; technology emulation and information migration.

As a first step in developing a set of tools to guide decision processes by digital heritage managers a study was commissioned to develop a framework for digital preservation issues.

As a second step along the same path this Cimtech study was commissioned to develop the next level of decision tools. Three aims were set for the study:

- to draw up a matrix of data types and categories of digital resources
- to draw up a decision model for assessing the agreed categories of digital resources to determine the most appropriate method of long term preservation
- to draw up a cost model for comparing the costs of the preferred methods of preservation for each category of digital resource.

1.3 Study Methodology

The consultancy drew heavily on work done in the same programme of studies and worldwide to draw up and agree a framework within which collection managers, or "caretakers" as they are referred to above, can address issues and develop data policies.

The consultancy reviewed existing work to develop cost models both in the UK and worldwide.

The consultancy carried out a wide-ranging review of the literature. The resulting bibliography is attached as an appendix to the report.

The consultancy visited a number of both well-established and newly formed digital libraries and data centres.

The consultancy defined the scope of the project by referencing a framework for digital collection management; by defining what is meant by digital preservation; by defining the key digital preservation strategies and by defining the range of data types and categories of digital resources covered. The definitions are provided in chapter 2 and the data types and digital resource categories are listed in chapter 3.

The consultancy then defined the required decision model and cost model and applied these models to the categories of digital resources defined in the study. These are presented in Chapter 4 and 5.

2 Defining the Scope Of The Study

The terms of reference for the study pointed out that the number of digital documents and resources to be managed and preserved is growing at a tremendous rate. Almost inevitably, the number of approaches being taken to the management and preservation of digital documents and resources is also growing rapidly. This raised four important questions that had to be addressed at the outset in order to build on the terms of reference and agree the scope and boundaries of the study.

2.1 Defining A Framework

The first question relates to the intended audience for this study. Who are the collection managers or caretakers and what are they trying to do? Are librarians, records managers, archivists and the managers of data centres all trying to do the same things with digital resources now – and in the future? If not then where do the similarities and the differences lie? What impact do these differences have on their preservation requirements?

In order to answer these questions the consultancy needed to identify an overall framework. An excellent framework has been developed and described by Daniel Greenstein (3). This is aimed at managers of collections of digitised scholarly and cultural information but is of even wider applicability. His objective in proposing the framework was to assist collection managers in identifying and addressing key issues and in developing their own data policies.

The consultancy's objective in drawing on Greenstein's proposed framework in this study is to define and agree the context within which digital preservation is being addressed. Preservation is one of seven modules that together comprise Greenstein's proposed framework. The seven modules described by Greenstein are as follows. Some of the key modules are further subdivided.

	Module name	Sub-module name	Sub-sub-module name
1	Data Creation		
2	Data Selection & Evaluation		
3	Data Management	Data Structure	
		Data Documentation	
		Data Storage	
		Data Validation	Data Assessment
			Data Copying
			Media Refreshment
4	Resource Disclosure		
5	Data Use		
6	Data Preservation		
7	Rights Management		

- **Data creation** (decisions made when the digital resource was created – often outside the control of the collection manager but having a major impact on the options subsequently open to the collection manager)

- **Data selection and evaluation** (decisions based on the digital resource's content, usability and relevance to the user base plus, on the ease with which the digital resource can be managed, catalogued, made accessible to users and preserved)

- **Data management** (broken down into four subsets)

  Data structure (how a digital resource is formatted, compressed and encoded which determines whether the digital resource needs to be re-formatted, uncompressed and unencoded or re-encoded)

  Data documentation (the extent to which the digital resource's structure, content, provenance and history have been documented – which may determine whether the resource needs additional documentation)

  Data storage (the computer hardware and media used to store the digital resource; whether the digital resource is stored online or offline and whether the storage is provided in-house or by a third party)

  Data validation (three procedures designed to ensure a digital resource's integrity)

    Data assessment (testing the digital resource's completeness; function and consistency)

    Data copying (making additional copies of the digital resource to guard against the loss or corruption of any one copy)

    Media refreshment (periodically copying one copy of a digital resource onto fresh media to protect against the corruption and content loss which may result from media deterioration)

- **Resource disclosure** (making information about a digital resource available to users e g via online catalogues)

- **Data use** (how digital resources are to be delivered to end users and used by end users – will be influenced by how digital resources were created and managed and will influence how digital resources are managed)

- **Data preservation** (safeguarding the information content of any digital resource from the ravages of time, technological change and decaying magnetic media – different preservation strategies are appropriate for different data types and structures. Preservation requirements will impinge on how digital resources are structured, documented, stored and validated and possibly even on the conditions and methods by which digital resources can be accessed by end users)

- **Rights management** (intellectual property rights, data protection and confidentiality issues – need to develop both acquisition licenses and distribution licenses and implementation procedures)

Greenstein goes on to make the vital point that, in order to implement the framework in any one situation, the collection manager or caretaker needs to turn it into a data policy that suits their particular collection's needs. The needs of librarians and records managers and data centre managers will vary but they can all be described and differentiated within this framework.

Greenstein further makes the point that each collection manager needs a clear understanding of the collection's aims and its overall budget. In all cases cost will be a vital consideration. This is particularly true when it comes to the area of preservation. As we shall see below – some preservation strategies today require almost unlimited resources and hence are not practical for the majority of collection managers.

The last point that Greenstein makes which is vital to this study - is that all the key collection management issues and hence all the seven modules of the framework are closely inter-related. Decisions about whether to create or include a digital resource in a collection - and about its content and format – will impinge on how it can be managed and stored, on how or even whether it can be preserved and on how it can be delivered to end users. Equally, the uses intended for a particular resource or the method chosen to preserve it over time may impinge upon decisions taken when creating or including a digital resource into a collection.

The close inter-relations and dependencies between all aspects of digital collection management make it very difficult, in practice, to identify and isolate those actions that simply relate to preservation. There are some tasks that specifically relate to preservation but a successful preservation strategy depends upon good practice in almost all the other six areas identified in the framework.

For the same reason it has proved difficult to isolate just that subset of a data centre's budget which relates to preservation. Some costs specifically relate to preservation but one cannot infer that those are the only costs incurred when preserving digital resources. Given that a successful preservation strategy depends upon good practice in all the other six areas in the framework one also needs to allow for the costs incurred in each of those other six areas. This is the approach that has been adopted by the consultancy when drawing up the cost model.

2.2 Defining Preservation

The second question relates to what is actually meant by the term "preservation" in this context – why we need to preserve digital resources and what is involved in actually "preserving" digital resources.

Jeff Rothenberg (4) provides one of the clearest accounts of some of the key challenges and tasks involved in preserving digital resources. The consultancy draws heavily upon his excellent article in this subsection. Digital preservation is broken down into three areas, each of which can then be further subdivided.

2.2.1 Preserving Bit Streams through Copying/Refreshing

Digital resources can be stored on any medium that can represent their binary digits. Rothenberg defines a "bit stream" as "an intended meaningful sequence of bits with no intervening spaces, punctuation or formatting". To preserve that bit stream the first requirement is to ensure that the bit stream is stored on a stable medium. If the digital medium deteriorates or becomes obsolete before we have read the digital information off the medium and copied it onto another medium then we have lost the data.

At a basic level, therefore, digital preservation involves the following tasks:

- preserving the digital medium that holds the digital information by storing it in the correct environment and following agreed storage and handling procedures;

- copying the digital information onto newer, fresher media before the old media deteriorates past the point where the digital information can be read or becomes so obsolete that we can no longer find a storage device to read it on;

- preserving the integrity of the digital information during the copying process.

Preserving the integrity of the digital information at this level means preserving the bit configuration that uniquely defines the digital object. The Task Force on Archiving of Digital Information (1) confirms that "there are various well-established techniques, such as checksums and digests, for tracking the bit-level equivalence of digital objects and ensuring that a preserved object is identical to the original".

However, simply preserving the digital information on several copies of a stable, digital medium is not sufficient. We also need to be sure that the digital information can be retrieved and processed in future.

Retrieving a bit stream requires a hardware device such as a disk drive or a tape drive and special circuitry for reading the physical representation of the bits from the medium. Accessing the device from a given computer also requires a driver program.

So if we hold digital resources on a specific type of digital medium (CD-ROM; tape; diskette; punched card etc) we need a drive designed to accept that specific type of digital medium in order to read the data.

Today it is becoming difficult to find computer punch card reader suppliers so if you have not copied your old digital resources off punch cards and onto a newer media you will find it progressively more difficult to read them. The same problem could be experienced in future when CD-ROM drives or drives for specific formats of magnetic tape media become obsolete. This approach is referred to as copying or "refreshing".

In some cases, due to the way the digital resource has been recorded onto a specific type of digital medium, it can be difficult to retrieve the digital resource and write it onto a different medium. This is often the case with digital publications - digital resources that have been authored and published on a specific digital medium and are provided with proprietary access software to provide access to the data held in the publication.

The publisher may have set a limit to how much data could be copied or down loaded from the publication in one session. It may be impossible to copy the data using normal utilities without obtaining a key from the publisher. If the publisher cannot be contacted then it may prove impossible to copy or migrate data from the specific medium and hence no refreshing can take place. When the medium deteriorates or becomes obsolete the data will be lost.

A less extreme case would be where the data was authored and the access software was designed specifically for one type of digital medium. There would be a danger that if the data was copied to another type of medium - then all the links between the data and the access software used to retrieve the data would be lost and it might not be possible to use the access software at all. In such cases new access software would have to be written.

Such digital resources can be referred to as hardware – specific. Once the drives needed to read the media become obsolete and unusable it may prove impossible to access the digital publications in their original format using the original access software.

2.2.2 Ensuring We Can Interpret Data By Preserving Documentation

Rothenberg points out that – assuming you can physically read the bit stream – the next step is to be able to interpret it. This is not simple as a given bit stream (see section 3 below) can represent almost anything from a sequence of integers to an array of dots in an image etc.

Also, interpreting a bit stream depends on understanding its implicit structure which cannot be explicitly represented in the stream. A bit stream that represents a sequence of alphabetic characters may consist of fixed length chunks (bytes) each representing a code for a single character. To extract the bytes from the bit stream, thereby parsing the stream into its components, we must know the length of a byte.

One way to convey the length is to encode a key at the beginning of the bit stream but Rothenberg (4) points out that this key must itself be represented by a byte of some length. A reader needs another key to understand the first one. The solution to this recursive problem is traditionally a "bootstrap" in the form of some human readable documentation which explains how to interpret the digital resource.

After the bit stream has been correctly parsed the next recursive problem involves interpreting the bytes. A byte can represent a number or an alphabetic character, according to a code. To interpret such bytes we need to know the coding scheme and the solution again is documentation that explains the byte encoding scheme. Hence, in addition to preserving the digital resource on a currently readable digital storage medium, we need to preserve documentation that allows us to interpret the digital resource.

Rothenberg goes on to point out that bit streams are usually stored as a collection or file of bits that contains logically related but physically separated elements. These elements are linked to each other by internal references consisting of pointers to other elements or of patterns to be matched.

Hence, in addition to simply reading the bit stream, there is a requirement to be able to interpret the information embedded in the bit stream. Most files contain information that is only meaningful to the software that created them. Word Processing files embed format instructions describing typography, layout and structure. Spreadsheet files embed formulas relating to their cells etc. This embedded information and all aspects of the representation of a bit stream - including the byte length, character code and structure - comprise the encoding of a file. The files contain both instructions and data that can only be interpreted by the appropriate software.

2.2.3 Options For Ensuring We Can Decode Data In Future

A word processing file does not represent a document in its own right. It merely describes a document that comes into existence when the file is interpreted by the program that produced it. In a very telling phrase, Rothenberg makes the point that "without this program or equivalent software, the document is a cryptic hostage of its own encoding".

To “preserve” a digital resource we need to ensure that we can decode the digital resource in future. There are two main approaches taken to solving this complex requirement.

The conservative approach assumes that the only way to ensure that you will be able to fully decode the bit streams held in a file in future is to preserve the program used to create it. If you have documents held in “Microsoft Word version 6.0” format then you also need to preserve “Microsoft Word version 6.0”. This approach is reviewed below as option one – preserving the original program. It leads to one of two preservation strategies – “technology preservation” or “technology emulation”.

The optimistic approach is that the best way to ensure you will be able to fully decode the bit streams held in a file in future is to ensure that they are encoded in a format that is independent of the particular hardware and software used to create them. It is vital with this approach to also ensure that there is always software available to decode the current format. This approach is reviewed below as option two. It leads to one preservation strategy – “digital information migration”.

2.2.3.1 Option One - Preserve the Original Program

Rothenberg is a strong advocate of this approach. He concedes that you do not always need to run the specific program that created a document in order to read that document. If it is a simple document then similar software may be able to, at least partially, interpret the file. But Rothenberg feels it is naïve to expect that the encoding of any new digital document will remain readable by future software for very long.

His view is that IT creates new schemes that often abandon their predecessors instead of subsuming them. The latest version of the leading word processing package should be backward compatible with the previous one or two versions. However, in his view, it would be naïve to expect all future versions for the next ten years to be able to read all the files created on all the oldest versions of the package.

Where we are dealing with very sophisticated digital resources such as multimedia presentations etc - there may be a total dependency between the digital resource and one version of one software package. Such a digital resource is therefore said to be “software-dependent”.

Often - where some level of interchange is permitted between programs - it will involve some loss of data. 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). However, Rothenberg points out that authors rarely save their work as pure text. They want to store format data and figures and footnotes. If this data is lost in the interchange then valuable content is lost.

Rothenberg therefore argues that, if a scholar or researcher wants to view a complicated document as its author created it, he or she may have no choice but to run the application software that was used to create it. Hence in order to fully decode the bit streams held in a file and to view a digital resource such as a complicated document in its original format then we may also need to preserve the program used to create it.

If we archive the original program with the files it created then in future, when both are retrieved, there will also be a need to find and use the operating system software which the original program was designed to run on. Depending on how long the data and the original program are kept for then this can become a serious problem.

The logical next step with this approach is, therefore, to plan to archive a copy of the operating system software with all the original application programs that run on it and all the files to be decoded by those application programs. Depending on how long the data and the application program and the operating system software are kept for, the final problem that occurs with this approach can then be finding the hardware on which to run the operating system software. There are only two solutions proposed to this problem.

The first is to preserve working replicas of all key computer hardware platforms. This is referred to as the "Technology Preservation" strategy.

The second is to program future powerful computer systems to emulate older obsolete computer platforms and operating systems on demand. Your latest PC server could be programmed to emulate a specific VAX computer running a specific version of the VMS operating system etc. This is referred to as the "Technology Emulation" strategy.

Both of these two potential digital preservation strategies are reviewed briefly below.

2.2.3.2 Option Two - Digital Information Migration

This approach is the opposite of option one above. It assumes that the best way to ensure you will be able to fully decode the bit streams held in a file in future is to ensure that they are encoded in a format that is independent of the particular hardware and software used to create them. It is vital with this approach to also ensure that there is always software available to decode the current format.

Digital information migration could potentially be facilitated in the following ways:

- Through copying the digital information to an analogue medium
- Through application programs that are "backward compatible"
- Through application programs that can "interoperate" with competing products
- Through converting digital resources into a small number of "standard" formats that are hardware and software independent

Digital information migration is the third digital preservation strategy covered by this study. It is reviewed in detail below.

2.3 Defining the Digital Preservation Strategies

The third question relates to the range of potential digital preservation strategies that should be covered in the study and a definition and assessment of each one. The terms of reference for the study calls upon the study to cover three potential strategies for ensuring long term access to digital information.

The three strategies were listed as being:

- Technology preservation
- Technology emulation
- Digital information migration

Section (2.2) above defined what we mean by digital preservation and placed these three strategies in the context of the two differing approaches that have been taken to digital preservation. It is rarely the case that a collection manager would want to adopt all three strategies for the preservation of one category of digital resources.

Normally the collection manager would select one or other strategy as being the most appropriate for each category of digital resource. However, it is quite possible that within one large collection there may be a requirement to adopt two or even three of the strategies for the long term preservation of a range of different categories of digital resources.

This explains the need for the first decision model. It aims to provide collection managers with guidance on the choice of the most appropriate long term preservation strategy for each of the different categories of digital resources.

The three strategies are described in more detail below. They are related back to the preservation requirements defined in section (2.2) above. The consultancy refers to these definitions later in chapters four and five of the study.

2.3.1 Technology Preservation

This strategy involves the following tasks:

- storing the bit streams on a stable digital medium;
- preserving the digital medium while the bit streams are stored on it;
- refreshing or copying the data to new media as required;
- preserving the integrity of the digital information during the copying process;
- preserving the original application program used to create or access the digital resource;
- preserving the operating system software that the original application programs run on;
- preserving the computer hardware platform that the operating system software was designed to run on.

The advocates of this strategy stress that to really replicate the behaviour of a program and the look and feel of a document or publication then you need to be running the original environment. While this is undoubtedly true in a purist sense it 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. Examples of some of the more popular old platforms are still kept going by some computer enthusiasts as a hobby but they are fighting a losing battle trying to source old components.

Today we are seeing the increasing dominance and ubiquity of a few computer platforms. In theory this should simplify the task of preserving these platforms 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 cannot be regarded as viable for anything other than the short to medium term. The consultancy would see "technology preservation" being used as a relatively desperate measure in cases where valuable digital resources cannot be converted into 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.

This strategy could be adopted where the only way to access a valuable digital resource was via an application that would only run on operating system software that would itself only run on an obsolete hardware platform. In this situation then collection managers would be best advised to seek out a specialist third party (if one could be found) with that hardware environment. They should then run the software and attempt to migrate the data off to at least a software-dependent format and ideally to a software independent format, which they can then migrate forward.

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.

2.3.2 Technology Emulation

This strategy has a lot in common with the technology preservation strategy described above. It involves the following tasks:

- storing the bit streams on a stable digital medium;
- preserving the digital medium while the bit streams are stored on it;
- refreshing or copying the data to new media as required;
- preserving the integrity of the digital information during the copying process.
- preserving the original application program used to create or access the digital resource.

Where it differs from the previous strategy is in how it creates the operating system and hardware environment that the original application program was designed to run on. This strategy does not involve preserving ageing hardware and running the original operating system software on top of it. Rather, it involves software engineers performing the following tasks:

- designing and running 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.

In other words you could configure your future PCs to look like a specific model of a VAX computer running a specific version of the VMS operating system etc. Rothenberg advocates this approach but concedes that it will require extremely detailed specifications for the outdated hardware and operating system software.

Looking to the future, one can agree with Rothenberg that some applications and operating system software such as MS DOS may remain ubiquitous so that all a collection manager would need to do would be to refer users to these programs.

Also, when these ubiquitous proprietary programs become obsolete and hence less commercially valuable, then copyright restrictions tend to expire and they can stay available to future users.

In general, the consultancy would still see this as a short to medium term strategy or a specialist strategy where the need to maintain the look and feel of the original digital resource is of great importance to the collection's user base.

The consultancy would see technology emulation being primarily used in cases where digital resources cannot be converted into 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.

This strategy could be adopted where the only way to access a valuable digital resource was via an application that would only run on operating system software that would itself only run on an obsolete hardware platform.

In this situation then collection managers would be best advised to seek out a specialist third party (if one could be found) able to emulate that hardware and operating system software environment. They should then run the software and attempt to migrate the data off to at least a software-dependent format and ideally to a software independent format so they can then migrate the data forwards.

Any collection manager in charge of a large collection of digital resources who relied solely on this strategy currently would be taking a very significant risk. They would be depending on the technical ability of the software engineers to emulate a specific environment and sustain it and the commercial viability of anyone providing such a service.

2.3.3 Digital Information Migration

This strategy assumes that the best way to ensure you will be able to fully decode the bit streams held in a file in future is to ensure that they are encoded in a format that is independent of the particular hardware and software used to create them. It is vital to ensure that there is always software available to decode the current format.

Another way of paraphrasing this strategy is to say 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 the collection managers must migrate their digital resources forwards to ensure the digital resources can also be accessed on the new platforms.

The Task Force on Archiving of Digital Information (1) have provided a useful definition of digital information migration as follows:

"Migration is the periodic transfer of digital materials from one hardware/software configuration to another, or from one generation of computer technology to a subsequent generation. The purpose of migration is to preserve the integrity of digital objects and to retain the ability for clients to retrieve, display, and otherwise use them in the face of constantly changing technology. Migration includes refreshing as a means of digital preservation. It differs from it in the sense that it is not always possible to make an exact digital copy or replica of a database or other information object as hardware and software change and still maintain the compatibility of the object with the new generation of technology. Even for information that is encoded in a contemporary standard form (a bibliographic database in USMARC or a corporate financial database in SQL relational tables), forward migration of the information to a new standard or application program is, as anyone knows who has witnessed or participated in such a process, time-consuming, costly and much more complex than simple refreshing".

The Task Force goes on to make the point that given the fact that the computer operating environments of digital archives will inevitably change over time, migration is not optional – it is an essential operation. The Task Force makes the point that:

"There are a variety of migration strategies for transferring digital information from obsolete systems to current hardware and software systems so that the information remains accessible and usable. No single strategy applies to all formats of digital information and none of the current preservation methods is entirely satisfactory. Migration strategies and their associated costs vary in different application environments, for different formats of digital materials, and for preserving different degrees of computation, display, and retrieval capabilities".

2.3.3.1 Change Media

One basic migration strategy that the Task Force covers involves the transfer of digital resources from less stable to more stable media. They point out that the most prevalent version of this strategy involves printing digital information onto paper or recording it on microfilm. Paper and microfilm are more stable than most digital media and no special hardware or software are needed to retrieve information from them.

This strategy is, strictly speaking, outside the scope of this study. However, it is worth referring to briefly. It may be appropriate in those cases where collection managers are faced with having to preserve hardware and/or software – dependent digital resources for long periods of time with relatively low budgets.

The alternative strategies – technology preservation and technology emulation – are considered to be short to medium term strategies and are potentially extremely expensive. Hence, provided the digital resources are "document like" - they can be printed out onto paper or microfilm in a linear fashion – then printing or recording to microfilm should be considered. If nothing else it preserves a copy of the basic data while other migration strategies are explored.

The Task Force (1) comment that "copying from one medium to another has the distinct advantage of being universally available and easy to implement. It is a cost-effective strategy for preserving digital information in those cases where retaining the content is paramount, but display, indexing, and computational characteristics are not critical. As long as the preservation community lacks more robust and cost-effective migration strategies, printing to paper or film will remain the preferred method of storage for many institutions and for certain formats of digital information".

The Task Force go on to point out the disadvantages of this conservative strategy as well by stressing just how much valuable data may be lost by a decision to copy digital resources to paper or microfilm and rely exclusively on the paper or microfilm copies in future.

"the simplicity and universality of copying as a migration strategy may come at the expense of great losses in the form or structure of digital information. When the access method for some non-standard data changes, one must, in order to migrate them, often eliminate, or "flatten," the structure of documents, the data relationships embedded in databases, and the means of authentication which are managed and interpreted through software. Computation capabilities, graphic display, indexing, and other features may also be lost, leaving behind the skeletal remnant of the original object. This strategy is not feasible, however, for preserving complex data objects from complex systems. It is not possible, for example 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 back-up strategy or as a last resort if no other strategy meets the requirements.

2.3.3.2 Backward Compatibility

A second migration strategy relies on popular application software being "backward compatible". The latest versions of most popular word processing packages will be capable of decoding files created on earlier versions of the same package – particularly the previous two or three versions. If the leading application packages are "backward compatible" then migration simply involves testing the process and then loading files created on previous versions of the application program into the new version and saving them in the new file format.

While this strategy may work over the short term for simple digital resources created on some of the leading application packages it cannot be relied upon over the medium to 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 and hence be unable to support their software. They may be forced to introduce a totally new software package and drop support for their old package. They may be forced to redesign the package to stay competitive with the result that the new version cannot read old files created on earlier versions.

Hence any digital information migration strategy which simply relied on "backward compatibility" of the leading application software packages would represent a short-term strategy that exposed the collection to many risks.

2.3.3.3 Interoperability

The third migration strategy relies on "interoperability" between rival popular application programs. You do not always need to run the specific program that created a digital resource in order to read that digital resource. Digital resources created on one application program can be exported in a common interchange format and then imported into a rival application program.

If such “interoperability” could be guaranteed between all the major competing application programs then “digital information migration” would be a much easier process. If your preferred CAD supplier ceased trading you could export all your CAD designs in a common interchange format and then import them into your preferred new CAD application. You could export all your alphanumeric data from your old database engine and load them into your new preferred database engine. You could export all your text files from your old word processing application and load them into your preferred new word processing application. You could export all your raster image files from your old image processing application and load them into your new image processing application. The list of migration options would be endless.

If it is a simple digital resource then the fact is that, today, similar software may well be able to, at least partially, interpret the file. However, it is common where some level of interchange is permitted between programs, for the interchange to involve some loss of data. 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). However, as Rothenberg points out (4), authors rarely save their work as pure text. They want to store format data and figures and footnotes. If this data is lost in the interchange then valuable content is lost.

The more complex the digital resources then the more problems are involved in “interchange” and the more valuable data is likely to be lost in the process. How important the “loss of data” is will vary depending on the type of digital resource, the objectives of the collection manager and the needs of the users of the collection. Compared to all the data that is lost when digital resources are printed out to paper or microfilm, the data lost during such an “interchange” may be minor. On the other hand – when interchanging the data held in complex databases such as GIS databases and groupware databases – it could involve the loss of thousands of links that have taken years of effort to create and that represent the bulk of the value of the database.

While this strategy may prove useful over the short term - for migrating simple digital resources out of obsolete application packages and into preferred application packages - it cannot be relied upon over the medium to 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 interoperability between their own and rival application packages. They may go out of business and hence be unable to support their software. They may be forced to introduce a totally new software package and drop support for their export and import options. They may be forced to redesign the package to stay competitive with the result that the new version cannot read old files created on rival packages. The interchange formats themselves may cease to be supported or may be replaced by newer, richer formats.

Hence any digital information migration strategy which simply relied on “interoperability” between the leading application software packages would represent a short-term strategy that exposed the collection to many risks.

2.3.3.4 Conversion to Standard Formats

A fourth and final migration strategy was again covered by the Task Force (1). They suggest that it is particularly appropriate for digital archives with large, complex, and diverse collections of digital materials. The proposed strategy is:

"to migrate digital objects from the great multiplicity of formats used to create digital materials to a smaller, more manageable number of standard formats that can still encode the complexity of structure and form of the original. A digital archive might accept textual documents in several commonly available commercial word processing formats or require that documents conform to standards like SGML (ISO 8879). Databases might be stored in one of several common relational database management systems, while images would conform to a tagged image file format and standard compression algorithms (e.g., JFIF/JPEG)".

This represents an enhanced version of the "interoperability" strategy. That strategy tends to rely on interchange formats that can be generated automatically from within the applications. Inevitably that tends to result in simple formats that involve the loss of significant data.

With this strategy the onus is placed on the collection manager to define the preferred formats and select the formats that are most appropriate for the digital resources they collect and the users they serve. The Task Force point out some of the benefits of this strategy over a simple "change media" approach:

"Changing format as a migration strategy has the advantage of preserving more of the display, dissemination, and computational characteristics of the original object, while reducing the large variety of customised transformations that would otherwise be necessary to migrate material to future generations of technology. This strategy rests on the assumption that software products, which are either compliant with widely adopted standards or are widely dispersed in the marketplace, are less volatile than the software market as a whole. Also, most common commercial products provide utilities for upward migration and for swapping documents, databases, and more complex objects between software systems. Nevertheless, software and standards continue to evolve so this strategy simplifies but does not eliminate the need for periodic migration or the need for analysis of the potential effects of such migration on the integrity of the digital object".

The handling of text files or documents can provide us with one example of the difference between the "interoperability" and the "conversion to standard format" strategies. The interchange of basic text content has largely been solved today although the Task Force points out that there are still some issues that can effect interoperability:

"Text today is generally covered by a formal, international ASCII standard for representing character formats. Standard extensions exist for encoding diacritic characters in romance languages other than English and a new standard is slowly emerging to incorporate scripted languages under a new common encoding scheme (UNICODE). Alternative encoding methods, however, abound. IBM maintains its own EBCDIC character encoding scheme and Apple and Intel-based personal computers differ in the ways that they support extended ASCII character sets. For documents in which any of these character codes can adequately represent the contents, the differences in encoding schemes may matter little and digital archives can manage object integrity by mapping character sets from one to the other. However, for works involving multiple languages or complex equations and formula, where character mapping is imperfect or not possible, character set format takes on considerably more significance over the long-term as a matter of content integrity".

If we move beyond basic text content to cover text layout and structure coding then the differences between the two strategies become more apparent. The consultancy quotes from the Task Force report again:

"In addition to character set issues, digital archives must also grapple with the means of representing and preserving textual content embedded in layout and structure. Mark-up systems, such as implementations of TeX and the Standard Generalised Mark-up Language (SGML), do exist as platform-independent mechanisms for identifying and tagging for subsequent layout and retrieval detailed structural elements of documents. The use of TeX and its variants, for example, is relatively common among scholars in some scientific disciplines such as mathematics and computer science, and the federal government and scholarly publishers are increasingly employing the SGML standard in documents that they produce and distribute electronically. Beyond these relatively specialised segments, however, word processing and desktop publishing systems still dominate the market for the creation of documents with complex structure and layout, and the software for such use typically models and stores document structure and layout in proprietary terms. Although software may provide mechanisms for converting documents to common interchange formats, use of such mechanisms often results in the loss or inadequate rendering of content such as page structures and the layout of headers, footers and section headings".

So if we store documents in their native format then every time we interchange them we stand to lose valuable content. If, on the other hand the documents are marked up using agreed standards when they are created or when they are taken into digital collections then not only the basic text content but also the layout and structure of the documents can be preserved.

A Selecting A Format

Decisions on which formats to convert digital resources into should be based on the structure of the digital resources themselves; on the objectives set by the collection manager and on the requirements of the users of that collection.

One important issue relates to whether the top priority is given to preserving the ability to process or edit the digital resource (editability) or to preserving the format or visual presentation of the digital resource (format).

In most data centres one of the key objectives is to preserve the data and make it available in a format which allows it to be loaded into user application programs so it can be processed and new data derived from it. The data is the valuable resource and how that data is presented in the form of tables, graphs, and models etc is of secondary importance. In archives and records centres, on the other hand, the key objective may well be to preserve the format or visual presentation of the digital resource – to ensure its archival integrity.

In the latter case then one valid migration strategy for all electronic "document – like" records would be to convert them into a formatted page image format. Examples would include converting each electronic document into a series of PostScript page print files or to Adobe Acrobat PDF (Portable Document Format) files or to a series of page images stored in a raster graphics format. This strategy could be seen to be the digital equivalent of printing the digital resources out onto paper or microfilm. As with that strategy – so this simple strategy would not cater for time-based data (audio and motion video) and is far from ideal for three dimensional graphics and relational database records. This is looked at in more detail in subsection (3.4.2) below.

A slight modification of this strategy for all electronic "document-like" records would be to hold them in their native file format but store with them "file viewer" software that enables users to view and print the documents as formatted page files but does not allow them to edit the documents.

This latter strategy has some drawbacks as a “digital information migration” strategy. Firstly it would create a need to periodically migrate the “file viewer” software to ensure that the file viewer software could always operate in the current computer environment. In changing the preferred “file viewer” software the collection manager might find that the new “file viewer” cannot view some of the older files in the collection. In that case the collection manager has two choices.

Firstly, he/she has to go back and migrate the old files in which case we are back to selecting the right format to convert them to.

Secondly, he/she has to resort to a “technology preservation” or “technology emulation” strategy and store the old “file viewer” software and the operating system software it runs on and preserve the required hardware environment for the operating system software.

2.3.3.5 Summary

“Digital information migration” is not a simple agreed strategy. It shows the most promise for the future and is the most widely adopted strategy in all the centres contacted or visited by the consultancy. However, as the Task Force has pointed out, “No single strategy applies to all formats of digital information and none of the current preservation methods is entirely satisfactory”.

“Digital information migration” is evolving. Techniques for the migration of digital resources comprising simple files of data appear to be widely accepted and followed. On the other hand, to quote the Task Force once again "the preservation community is only beginning to address migration of more complex digital objects. Additional research on migration is needed to test the technical feasibility of various approaches to migration, determine the costs associated with these approaches, and establish benchmarks and best practices. Although migration should become more effective as the digital preservation community gains practical experience and learns how to select appropriate and effective methods, migration remains largely experimental and provides fertile ground for research and development efforts”.

2.4 Defining the Digital Resources Covered by the Study

The fourth question relates to the range of data types and the categories of digital resources that the study should cover and the influence they have on the choice of digital preservation strategy.

This is a problematic area given the sheer scale of the computing industry and the vast range of differing application programs on which digital resources are being created or managed. To produce an authoritative classification of all the data types; digital resource categories; application software; data structures and data management systems which are in current use or which have been used at some time in the past would be a lifetime's work.

A more pragmatic approach was required for this short study. The consultancy and their advisers, felt that it would be of most benefit to the widest audience for the study to concentrate on the basic data types which all but the most specialised data centres would need to collect.

Similarly the consultancy was advised to select a small set of common digital resource categories which all but the most specialised data centres would need to collect. The issues involved in preserving very specialised and very complex digital resources are only fully understood by the specialists who collect and manage them and the duration of the study did not allow the consultancy to seek out and research all the specialist data centres in the UK. The consultancy expects that the initial list of ten categories drawn up for this study would be modified and expanded in the light of further research. The aim would be to develop an agreed list of categories that could be referred to and used to commission more detailed studies into the best preservation strategy for each category in future.

2.4.1 Basic Data Types

The basic data types covered by this study are listed and described briefly in section (3.1) below. They form the building blocks from which all of the simple and complex categories of digital resources are constructed.

There are many application programs that still involve the processing of one basic data type and hence the creation of digital resources that comprise just one basic data type. Examples include the huge range of numeric and alphanumeric data processing applications; image processing applications and simple text processing applications.

However, in the increasingly sophisticated world of computing, many of our most popular application programs now involve the processing and creation of multiple data types and the creation of digital resources that comprise multiple data types.

2.4.2 Categories Of Digital Resource

The ten generic digital resource categories selected for this study are listed and described in section (3.2) below. It is far from being a definitive grouping but has been drawn up based on a review of the many lists and groupings discovered during the research for this study.

2.4.3 Application Programs

The key application programs that may be used to create the ten generic categories of digital resource (or the basic data types that comprise them) are reviewed in section (3.3) below. These would need to be preserved if a “technology preservation” or “technology emulation” strategy was adopted. They may also have an impact on the choice of migration strategy if a “digital information migration” strategy is adopted.

2.4.4 Structure Of Digital Resources

The structures that may have been used to store, interchange and present these ten categories of digital resource (or the basic data types that comprise them) are reviewed in section (3.4) below. They may have an impact on the choice of migration strategy if a “digital information migration” strategy is adopted.

2.4.5 Management/Distribution Systems

The systems that may have been used to compile/manage/distribute collections of the ten categories of digital resource are reviewed in section (3.5) below. These may need to be preserved if a “technology preservation” or “technology emulation” strategy was adopted. They may also have a significant impact on the choice of migration strategy if a “digital information migration” strategy is adopted.

3 Data Types & Digital Resources Covered by the Study

3.1 Basic Data Types Covered by the Study

	Data Type	Examples	Notes
1	Alphanumeric Data	Flat files; hierarchical and relational data sets; Office documents; marked up formal texts;	ASCII data plus proprietary or standard database codes; plus proprietary or standard mark-up codes
2	Raster Graphic Data	Bitonal, greyscale and colour images of pictures, documents, maps, photographs;	Standards for compression of image data and for interchange
3	Vector Graphic Data	Presentations; creative graphics, computer aided designs; clip art; line drawings; 3D models; maps	De facto interchange standards; standard graphics languages – GKS; PHIGS; PHIGS +
4	Sound/Audio Data	Speech processing; speech simulation; speech recognition; sound/music recording;	De facto audio file standards; sample rates, word sizes; symbolic music recording,
5	Motion Video Data	Moving images; full frame, full motion video; interleaved audio and video	Frame sizes, frame rates, resolution; raster graphics; standards for compression; capture or computer generation
6	Moving Graphics; Animation Data	Moving graphics; objects and timing relationships	Manual creation via animation authoring tools; Vector graphics; computer generation

3.2

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; presentation systems; Computer Based Training; multimedia	Business presentations; formal courseware; CBT packages;

	Category Of Digital Resource	Data Types Included In Resource	Applications Used To Create/Manage / Distribute Digital Resource	Notes
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; co-ordinates; range of overlay data; links between data types;
10	Interactive Multimedia Publications	Interleaved audio and video data; moving graphics; vector and raster graphics; alphanumeric data;	Authoring software; editing software; access software;	Electronic publishing; educational and training material; marketing material; games etc

3.3 Applications Used to Create the Digital Resource Categories

	Application Program	Data Type/s Created	Category Of Digital Resource Created	Notes
1	Data Processing; Environmental; scientific; finance; administration	Alphanumeric data	Data sets	Survey data; results of experiments; Transaction data; event data; administrative data; attribute/bibliographic data
2	Word Processing	Alphanumeric data; mark-up codes; graphic data; tables	Office documents; structured texts	Simple text documents; reports; literary texts; text for input to publishing systems; text for mark-up;
3	Desktop/ Corporate publishing	Alphanumeric data; mark-up codes; compound documents; tagged graphics; indexes;	Structured texts;	Reports; directories; catalogues; corporate publications; commercial publications
4	HTML Editors	Alphanumeric data; mark-up tags; Web pages	Structured texts	Simple documents; Web pages
5	SGML Editors	Alphanumeric data; mark up codes;	Structured Texts;	Reports; Corporate Publications; Commercial Publications
6	Web page design	Alphanumeric data; mark-up tags; raster /vector graphics	Web pages with graphics	HTML; JPEG & GIF
7	Spread-sheet packages	Alphanumeric data	Data sets/ Office documents	Can be stored in native form for access via spreadsheet or ASCII data can be extracted and held as data set; Data Interchange File (DIF)
8	Business graphics packages	Vector graphics; alphanumeric data;	Flow charts; line drawings;	Can produce graphics for import into compound documents
9	Creative graphics/ clip art	Vector graphics; raster graphics; alphanumeric data	Art work; advertising copy; clip art;	Can produce graphics for import into compound documents
10	Presentation graphics	Vector/raster graphics; alphanumeric data; moving graphics;	Presentations; Courseware; Slides	Adobe Persuasion; MS PowerPoint; Slide manager etc
11	Document Image Processing	Raster graphics;	Office documents; maps; designs; image collections	Used to capture digital images of paper documents; drawings etc. Form of image processing software; With recognition software can capture text from paper documents

	Application Program	Data Type/s Created	Category Of Digital Resource Created	Notes
12	Image editing/ processing	Raster graphics	Visual Images	Used to capture bitonal, greyscale or colour images of fine art, photographs etc and to manipulate and enhance the images
13	Computer Aided Design (CAD)	Vector graphics; alphanumeric data	Design Data; Mapping Data	Used to create 2 and 3 dimensional designs, models, plans etc
14	Simulation, Modelling & Testing	Vector graphics; raster graphics; animation; alphanumeric data	Ground modelling; flight simulation; 3D visualisation	Used to create 3D models and images and in simulation roles
15	GIS	Vector graphics; raster images; alphanumeric data; links	Mapping data; land cover; population trends	Used to create and manage (see 3.5) links between maps and overlaid data types e g Arc/Info; Arc/View MapInfo etc
16	Speech Processing	Speech coding; Speech synthesis; Speech recognition	Store and playback speech; Computer communication to humans; Create office documents; control computers	Speech recognition used to capture dictated text and load it into word processing packages for editing and to control computer systems
17	Music/ Audio Processing/ Recording	Audio data	Music recordings	Digital recording of live music or existing analogue recorded music; digital composition
18	Digital Video Recording; Processing Editing; Generation	Video data; audio data; interleaved audio and video	Video recordings	Digital recording of live video broadcast or existing analogue recorded video; computer generation
19	Animation Processing; Generation	Moving graphics; animation	Presentations; games; entertainment	Manual creation of moving graphics with authoring tools; computer generation of moving graphics

3.4 Structuring Data Types & Categories of Digital Resource

In section (3.1) the consultancy defined six data types which are created using a range of application programs.

The consultancy listed 19 examples of application programs in section (3.3) above. Many more could be defined and listed.

In section (3.2) the consultancy listed 10 categories of digital resources which are made up from one or more of the six data types.

In subsection (3.4.1) below the consultancy reviews how each of the six data types can be structured for interchange and preservation. This includes a list of content coding schemes and standards; compression algorithms used to compress the data types and file formats used to store the bit streams containing these data types.

In subsection (3.4.2) below the consultancy reviews how each of the ten categories of digital resources can be structured for interchange and preservation.

3.4.1 Structuring Data Types

	Data Type	Subset Of Data Type	Content Coding	Compression Algorithm	File Format
1	Alphanumeric data		ASCII		Comma Separate Variable (CSV); Delimited Field File (DFF); .TXT
			Extended ASCII		
			EBCDIC		
			Unicode		
		Spreadsheet Data			DIF (Data interchange file) spreadsheet, formula and data
2	Raster Graphic Data	Bitonal Image	1 bit per pixel	CCITT Group 3 facsimile; CCITT Group 4 facsimile	TIFF (tagged image file format)
		Bitonal Image	1 bit per pixel	JBIG	JBIG
		Bitonal Image	1 bit per pixel		GIF
		Bitonal Image	1 bit per pixel		PNTG MacPaint
		Bitonal Image	1 bit per pixel		PCX
		Base Greyscale	4 bits per pixel	JBIG	JBIG
		Full Greyscale	8 bits per pixel	LZW	TIFF
		Full Greyscale	8 bits per pixel	JPEG	JPEG
		Full Greyscale	8 bits per pixel	LZW	GIF
		Full Greyscale	8 bits per pixel	RLE	PCX
		Palette colour table	1,4,8 or 24 bits per pixel	RLE	Windows BMP

	Data Type	Subset Of Data Type	Content Coding	Compression Algorithm	File Format
2	Raster Graphic Data contd.	Palette colour table	1,4.8 or 24 bits per pixel	RLE	DIB (Device independent bitmap – Windows)
		Palette 256 Colours	8 bits per pixel	LZW	GIF
		Palette 256 Colours	8 bits per pixel	LZW	TIFF
		Palette 256 Colours	8 bits per pixel		Windows BMP
		Palette 256 Colours	8 bits per pixel	RLE	PCX
		Full Colour RGB	24 bits per pixel		TIFF
		Full Colour RGB	24 bits per pixel	JPEG	JPEG/JFIF
		Full Colour RGB	24 bits per pixel		Windows BMP
		Full Colour RGB	24 bits per pixel	RLE	PCX
		Full Colour CMYK	32 bits per pixel		TIFF
		Full Colour CMYK	32 bits per pixel	JPEG	JPEG/JFIF
	Raster and vector graphics	Full Colour			PostScript Level 1, 2 & Encapsulated (EPS)
	Raster and vector graphics				PICT vector and raster graphics format on Macintosh
	Raster and vector graphics		Used to interchange Word Processing graphics		WPG WordPerfect raster and vector graphics

	Data Type	Subset Of Data Type	Content Coding	Compression Algorithm	File Format
3	Vector Graphic Data	Two Dimensional	CAD and drawing programs		CGM Computer Graphics Metafile
		Two Dimensional	CAD and drawing programs		HPGL page description language with set of commands to describe graphics
		Two Dimensional			PIC (lotus 1-2-3 graphics)
		Two Dimensional			CDR Corel Draw vector formats
		Two & Three Dimensional	CAD		DXF AutoCAD vector format and de facto interchange format
		Two & Three Dimensional	CAD		DWG AutoCAD internal vector format
		Two & Three Dimensional			IGES Initial Graphics Exchange Specification (NIST)
4	Sound/Audio Data	Speech; Voice			
		Digitised sound signal	8 bit; 8 kHz sample rate (amplitude of each digitised sample in PCM G 711 digital telephony is represented with 8 bit code words)	G.721; G.722; G.723	ITU G.711 Pulse Code Modulation using logarithmic coding.

	Data Type	Subset Of Data Type	Content Coding	Compression Algorithm	File Format
4	Sound/Audio Data contd.	Sound			
		Digitised sound signal	CD Digital Audio; 16 bit 44.1 kHz sample rate (amplitude of each digitised sample represented with 16 bit code words		Pulse code modulation using linear coding
		Digitised sound signal	DAT; 12 & 16bit; 32, 44 and 48 kHz sample rates		
			MPC Level 1 8 bit word, mono		
			MPC Level 1 Sample at 11.025 kHz; play back at 11.025/ 22.05 kHz; 8 bit word		AU Sun format for UNIX adopted by Library of Congress
			MPC Level 2 Sample at 22.05 kHz; 16 bit word		WAVE (WAV) Microsoft format adopted by LC
			32, 44.1 or 48 kHz sampling rates;	MPEG-1; Audio Layer-1; Layer-2; Layer-3.	MPEG-1
		Symbolic coding	Musical Instrument Digital Interface MIDI FM synthesis or waveform synthesis		MID (notes + control change messages)

	Data Type	Subset Of Data Type	Content Coding	Compression Algorithm	File Format
5	Motion Video Data	Image size 160 x 120 pixels; 8 colours; frame rate 15 fps;		RLE; Intel Indeo	Microsoft AVI (Audio Video Interleaved)
		Standard Interchange Format SIF 352 x 240 pixels 30fps NTSC; 352 x 288 25fps PAL/ SECAM		MPEG – 1	MPEG - 1
		Low Level	CIF 352 x 288 samples/ frame	MPEG-2	MPEG-2
		Main Level ITU – R 601	720 x 480 samples/ frame	MPEG - 2	MPEG – 2
		High 1440 Level Consumer HDTV	1440 x 1152 samples/ frame	MPEG - 2	MPEG – 2
		High Level HDTV	1920 x 1080 samples/ frame	MPEG -2	MPEG – 2
		Very Low Bit Rate audio visual coding; video-conferencing	QCIF 10 fps 88 x 72 pixels	MPEG -4	MPEG – 4
		MPC-2 Image size 320 x 240 x 8 bit; 15 fps			
					Quick Time Video (Apple Format)
6	Moving Graphics/ Animation Data	Image size 160 x 120 pixels		RLE; Frame to frame differencing	Quick Time Animation (Apple format)
		Image size 320 x 200 pixels 8 bit		RLE; Frame to frame differencing	AutoDesk FLI/FLC Flick Formats

3.4.2 Structuring Digital Resources

In subsection (2.4.1) above the consultancy made the point that any one category of digital resource may comprise more than one basic data type. A digital resource is created by one or more application programs and can then be managed using one or more of the systems described in section (3.5) below. A "document" is the most common example of a digital resource and this section reviews some of the options available for structuring and interchanging digital documents.

A generic definition of a document is given by the consultancy in their "Document Management Directory" (5):

"a structured amount of information intended for human perception that can be interchanged as a unit between users and/or systems"

This is a very general definition of a document. It includes music and video, for example. The consultancy would also follow the pragmatic lead set by the Task Force (1) in distinguishing between two types of digital information resources that they describe as "document-like objects and other objects". To the Task Force:

"Document-like objects share the characteristics that they can, but need not be, adequately represented in print format. They include, for example, pure text, text with printable illustrations and photographs that can be recorded in print. Other objects cannot be represented in print and include sound, video, film, software and multimedia objects."

As defined, a document can be interchanged for either or both of the following purposes:

- To allow presentation as intended by the originator
- To allow processing such as editing and formatting

The composition of a document in interchange can take several forms depending on the purpose:

Formatted form	Allowing presentation of the document
Processable form	Allowing processing of the document
Formatted processable form	Allowing both presentation and processing

A document architecture can be defined (5) as being "rules for defining the structure of documents in terms of a set of components and content portions and the representation of documents in terms of constituents and attributes"

The structural information of a document consists of the set of one or more of the following structures: –

- specific logical structure;
- specific layout structure;
- generic logical structure;
- generic layout structure.

Layout is the physical viewpoint where a document may be seen as a collection of pages or images. The logical view sees the document in terms of its abstract components such as a collection of sentences etc.

A document has a specific and a generic structure. The specific document structure is the one that the user may read. The generic document structure is the template that guides the creation of the document and that could be re-used for its amendment.

The logical structure tends to be made up of some or all of the following logical objects:

- heading;
- table of contents;
- chapters; chapter headings;
- sections; section heading; section element;
- paragraphs,
- figures; picture, caption etc.

The layout structure tends to be made up of some or all of the following layout objects:

- pages
- blocks

One issue when drawing up a data management and digital preservation strategy is whether you are trying to preserve digital documents in a formatted form only or in a processable form from which you can derive formatted forms or in a formatted processable form.

If you scan and digitise the image/s of the pages of a paper document or the images of microfilm frames holding the image/s of the pages of a document then you are preserving digital images of the formatted form of that document.

The paper or printed version of a document represents the formatted form of the document.

If you store a digital document in the file/s created by the application program used to create it then you will be storing the digital document in a processable format from which you can derive formatted forms of the document by sending the document to print. The processable format will tend to be proprietary to the specific application program as we have seen above.

It may be possible to convert a digital document into an interchangeable processable format automatically from within the application. However, this will often result in loss of format data. Examples include converting text document into ASCII or RTF (Rich Text Format).

Certain categories of digital documents can be tagged and marked up and distributed or stored in a standard processable format. Examples would include text documents tagged using HTML and electronic texts marked up using SGML and an agreed Document Type Definition (DTD).

If you opt to preserve digital documents in a formatted form then a number of options are available. The Public Record Office's EROS (Electronic Records in Office Systems) programme has issued a valuable draft set of guidelines for the preservation of electronic records (6). They recommend the following formatted preservation formats.

The first is PostScript, a programming language for page layout and typesetting text and graphics on laser printers. Any application designed to run on desktop computers will support PostScript printing. PostScript was designed to be written to file as well as to a printer, which means that it can be used for electronic record transfer. The PRO like the fact that the use of PostScript requires little, if anything, by way of enhancement to existing IT applications.

The second is TIFF, which is recommended as a de facto standard for image. They would recommend it for the interchange of digital images.

The third is the Portable Document Format (PDF) developed and employed by Adobe in its Acrobat family of products. Digital documents can be converted into PDF format via PostScript. PDF is positioned as "electronic paper" for platform and application independent electronic record access and usage. The PRO point to the fact that the PDF specification has been placed in the public domain by Adobe.

	Category Of Digital Resource	Data Type/s	Proprietary Processable Forms	Standard Processable Forms	Standard Formatted Forms	Notes
1	Data Sets	Alphanumeric data		ASCII; CSV; Delimited	PDF Postscript	
2	Structured Texts	Alphanumeric data; mark-up data; tags to graphics;	WP Formats; DTP Formats;	SGML; HTML;	PostScript PDF TeX DSSSL	
3	Office Documents	Alphanumeric data; raster & vector graphics; Moving graphics	WP; Images Spreadsheets; Presentation Graphics;	ASCII; RTF; HTML; SGML; TIFF; CGM;	PostScript PDF; DSSSL	
4	Design Data	Vector/ raster graphics; alpha-numeric data	CAD formats;WP formats;	DXF/DWG; IGES; CGM; TIFF; ASCII/RTF	HP GL PostScript EPS	

	Category Of Digital Resource	Data Type/s	Proprietary Processable Forms	Standard Processable Forms	Standard Formatted Forms	Notes
5	Presentation Graphics	Vector/ raster graphics; alpha-numeric data; Moving graphics	Graphics formats; PowerPoint etc;		PostScript PDF	
6	Visual Images	Raster graphics	BMP; PCX;	TIFF; GIF; JPEG;	PostScript PDF	
7	Speech & Sound Recordings	Audio Data	Sun AU (UNIX) MS Wave	MPEG-1 Audio Layers 1/2/3 MIDI		
8	Video Recordings	Video Data	MS AVI; Apple Quick-time	MPEG-1 MPEG-2 MPEG – 4		
9	Geographic/ Mapping Data	Vector graphics; raster graphics; Alpha-numeric data	Arc/Info Arc/View MapInfo AutoCAD Map	TIFF; ASCII CGM	PostScript EPS HPGL	
10	Interactive Multimedia Publications	Audio/ video data moving graphics Raster/ vector graphics; alpha-numeric data	Macro-media; Apple Quick-time;	MPEG-1 MPEG-2		

3.5 Management/Distribution Systems

In the first four sections of this chapter the consultancy has defined ten generic categories of digital resource. The consultancy has described three technical factors that may impact on the choice of a preservation strategy for each category. These are the basic data types employed in each category; the application programs used to create each category and the structures applied to each category. This section covers the fourth factor that may influence the preservation strategy for each category. This relates to how the digital resources were managed and distributed/used prior to deposit or at the point of deposit.

In some cases, the digital resources will be live and the collection manager will be in contact with the creators prior to deposit. In these cases then how the digital resources have been managed and distributed in the past may be of little significance. The collection manager will be able to discuss the options with the creator/owner of the data and arrange for the digital resources to be processed and deposited/interchanged by the creators/owners in the most appropriate format. The only limiting factor may be the export options supported on the creator/owner's system.

In other cases, the digital resources will have been archived or published on a specific system and the creators and/or the system will no longer be available to process them. In these cases then the systems used to manage and/or distribute the digital resources will have a major impact on the preservation strategy adopted.

The system used to manage digital resources may manage a considerable volume of attribute or index data about the resources. It may have been used to hold thousands of links between data within different digital resources. Finally, it may have been used to provide users with powerful tools to navigate through the contents of the digital resources that it managed.

The creators/depositors and/or the collection manager would need to determine the importance of the attribute data; the links and the navigational aids. They would need to determine the resources involved in preserving all this additional data and they would need to agree the best approach to preserving the data.

The table below lists the ten categories of digital resources covered by this study and indicates the range of systems that may have been used to manage attribute data; links and navigational aids for each category. The second column covers the file management facilities available in operating system software. The third column covers the facilities for object linking and embedding supported by office application programs and suites. The fourth column covers the range of database management systems that have been used to manage ASCII data sets.

The fifth column covers the full range of object or content management systems designed to facilitate the creation and management of document databases. The sixth column covers the growing range of Web based systems designed to facilitate the creation and management of document databases on Web servers. The seventh column covers the range of CD publishing systems that can be used to assemble collections of digital resources on CD and provide links between them and a range of navigational aids. Many libraries are seeking to preserve such digital publications. The final column covers a range of interactive multimedia systems that are used to author and publish and provide interactive access to collections of multimedia objects on CD or online. Again, many libraries and archives are faced with trying to preserve such publications/databases.

Digital Resource	O S File Management Systems	Application/ Office Suite Links	ASCII data management	Object/ Content management systems	The Web	CD ROM Publishing	Interactive Multimedia Systems
Data sets			Flat File; Hierarchical; Relational Object		Distribution	Archive sets/tables	
Structured texts	Folder; File Name; Suffix Attributes;	OLE for graphics DTP	Attribute data in Flat File; Relational Databases	SGML databases; Proprietary compound document management systems	HTML; Gateway to SGML databases; Gateway to proprietary databases	Publishing and archive	
Office documents	Folder; File Name Suffix Attributes	OLE; Office bundles DTP	Attribute data in Flat File; Relational Databases	Proprietary document management systems; Groupware databases;	HTML; Gateway to proprietary document management databases	Archive	
Design data	Folder; File Name; Suffix Attributes	OLE	Attribute data in Flat File; Relational; Object Databases	Proprietary CAD and PDM/EDM databases	Gateway to proprietary PDM/EDM databases	Archive; Customer document-ation	
Presentation graphics	Folder; File Name; Suffix Attributes	OLE Office bundles	Attribute data in Flat File; Relational databases	Proprietary document management systems; Groupware databases	Gateway to proprietary document management databases	Archive; Publish CBT	
Visual images	Folder: File Name; Suffix; Attributes	OLE Office bundles; DTP	Attribute data in Flat File; Relational; Object Databases	Proprietary image databases; proprietary document management databases	Gateway to proprietary databases; HTML & SGML databases; PDF/GIF	Archive; Publish Photo CD	
Speech/ Sound Recordings	Folder; File Name; Suffix Attributes	OLE DTP	Attribute data in Flat Files; Relational; Object Databases	Proprietary document management systems	Distribution	Archive; Publish CD DA;	Audio & Video Interleaved
Video Recordings	Folder; File Name; Suffix Attributes	OLE Native	Attribute data in Flat Files; Relational; Object Databases	Proprietary document management systems	Distribution	Archive; Publish DVD; Video and Audio Interleaved	Author, Edit, Access Video & Audio Interleaved
Geographic/ Mapping Databases	Folder; File Name; Suffix Attributes	OLE Native Co-ordinates	Attribute data in Flat Files; Relational; Object databases	Proprietary GIS systems links/co-ordinates vector, text and ASCII data to bit maps	Gateway to proprietary GIS databases	Archive; Publish Gazeteers; Atlases; GIS databases	
Interactive Multimedia Publications	Folder; File Name; Suffix Attributes	OLE Native	Attribute data in Flat Files; Relational; Object Databases	Object Databases; Links; Shared Objects	Distribution	Archive; Publish MMPC	Collect, Publish, Link Clips, Navigate & Access

4 Developing a Decision Model

4.1 Selecting the Most Appropriate Long Term Preservation Strategy

As described in chapter 1 above, the terms of reference for this study called on the consultancy to "draw up a decision model for assessing the agreed categories of digital resources to determine the most appropriate method of long term preservation".

In chapter two the consultancy has defined what is meant by digital preservation and has defined the three main strategies adopted for digital preservation. The three main strategies and key subsets of the third strategy are as listed in the table below.

	Preservation Strategy	Subset of strategy
1	Technology Preservation	
2	Technology Emulation	
3	Digital Information Migration	Change Media
		Backward Compatibility
		Interoperability
		Conversion to standard formats

In chapter two the consultancy also referenced Greenstein's framework (3) which defines the context within which digital preservation is being addressed in this study. The framework defines seven modules, which must be addressed by any digital collection policy, of which preservation is one. The seven modules and key sub modules are as listed in the table below.

Module number	Module name	Sub-module name	Sub sub-module name
1	Data Creation		
2	Data selection and evaluation		
3	Data management	Data structure	
		Data documentation	
		Data storage	
		Data validation	Data assessment
			Data copying
			Media refreshment
4	Resource disclosure		
5	Data use		
6	Data preservation		
7	Rights management		

In chapter three the consultancy has defined 10 categories of digital resources and has reviewed the data types; applications; structures and management/distribution systems which can be employed to create, manage and distribute each category of digital resource.

The diagram (Fig 1) below, shows 7 key factors which need to be taken into account when deciding on the preferred preservation strategy for each category of digital resource.

However, in section (2.4) above, the consultancy has already reviewed the three preservation strategies and presented the third strategy – "digital information migration" – as being the preferred long - term strategy where it is technically and economically feasible. Migration was also the preferred long - term strategy of all the technical experts interviewed in the course of the study.

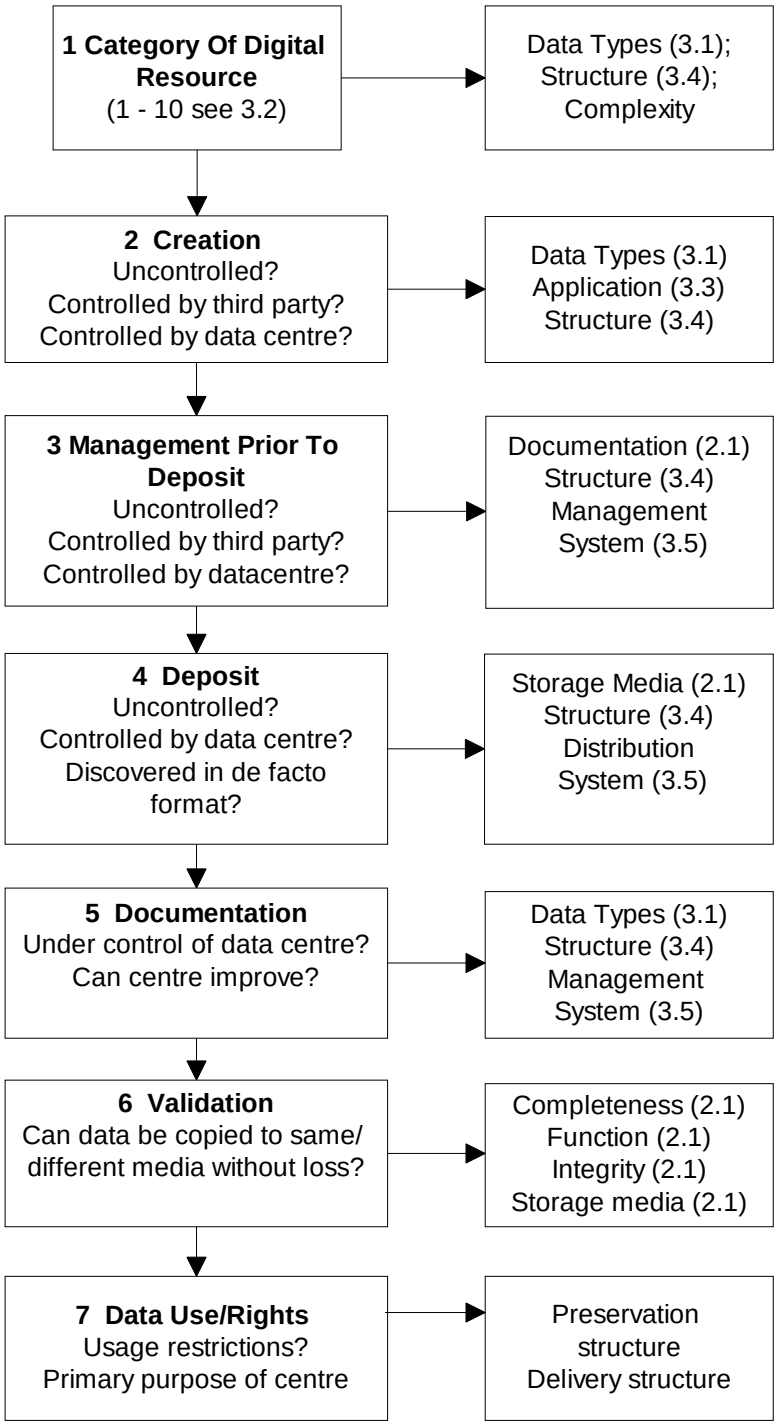
Given this level of consensus about the preferred preservation strategy, the consultancy proposes that, for each category of digital resource, the 7 key factors should be reviewed and used in the following two ways:

Firstly, to see whether any of them prevent the use of a "digital information migration" strategy. If one or more factors do appear to rule out the use of "digital information migration" for a specific category of digital resource then the other two strategies – "technology preservation" or "technology emulation" should be considered.

Secondly - if none of the 7 factors appear to rule out "digital information migration" as the preferred strategy – the 7 factors should be used to determine which subset of the strategy (change media; backward compatibility; interoperability or conversion to standard formats) is most appropriate for each category of digital resource. In some cases the review will indicate that there is a need to combine several of these subsets to form the perfect strategy for preserving a specific category of digital resource.

Fig 1

7 Factors Determining The Appropriate Preservation Strategy For Each Category Of Digital Resource



4.1.1 Category of Digital Resource

The first key factor relates to the basic data types employed in each category and the structures applied to each category.

4.1.1.1 Data Types

If the digital resource only contains one basic data type then this simplifies the management and preservation issues and helps to make “digital information migration” (migration) a more attractive and cost effective option. The more data types contained within a digital resource then the more complex the management and preservation issues become and the less attractive some of the “migration” subsets become.

If the digital resource only contains one basic data type and there are international or de facto standards for structuring and encoding that data type (ASCII for alphanumeric data) then this also makes the “interoperability” subset more attractive.

4.1.1.2 Data Structures

If the digital resource is held in a standard processable form then this can simplify the management and preservation issues. It helps to make “migration” an attractive option. “Interoperability” and “conversion to standard formats” may both be considered as feasible subsets of the “migration” strategy in this case.

If the digital resource can be said to be “document-like” (it can be printed out on paper as defined in (3.4.2) above) and can be exported in a standard formatted form then this can also simplify the management and preservation issues. It helps to make “migration” an attractive option. “Change media” and “conversion to standard formats” can both be considered as feasible subsets of the “migration” strategy in this case.

If the digital resource is held in a proprietary processable form and, as indicated below, it was created on an obscure application program which is no longer supported, then this can make the management and preservation issues very much more complex. “Migration” may be resource intensive and very costly in this case. If a copy of the original application program still exists then the collection manager may have to resort to one of the other, two digital preservation strategies – “technology preservation” or “technology emulation”.

4.1.2 Digital Resource Creation

The second key factor relates to how the digital resource was created and covers several areas. The first relates to the application program used to create the digital resource.

4.1.2.1 Application Programs

If the digital resource was created using one de facto standard application program that is widely supported then this can simplify the management and preservation issues. It helps to make “migration” an attractive and cost effective option. “Backward compatibility” and “interoperability” may be considered as feasible subsets of the “migration” strategy.

If the digital resource was created using an obscure application program which is no longer supported then this can make the management and preservation issues more complex.

4.1.2.2 Guidelines & Controls

The second relates to whether or not the digital resource was created in a controlled environment. The control could be imposed by the individual author or by the organisation he or she was working for or by the data centre or archive itself.

Historically, what controls there were would have been exercised by the authors themselves. If the author belonged to a profession then there may be standards followed for the production of formal interchange documents etc. Organisations may impose controls by mandating the use of standard templates and house styles.

Finally, the data centre or the archive may actually be able to set the standards. One example where this is beginning to happen is in the Research Councils. The Research Councils sponsor research by individuals and institutions. The same Research Councils also fund their own data centres to archive the results of these research projects. Increasingly the data centres are able to define standards for documenting the data and standards for structuring the data. The Research Councils then make it a condition of any research grants that the research data should be presented in the data centre's standard formats at the end of the project.

The Public Record Office is the ultimate archive for government records. Increasingly, as government departments create more data sets and electronic records, the PRO will also become the ultimate government data centre. The PRO is in a position to advise government departments on how best to implement electronic records management systems and how they should structure their electronic records for deposit. They have already produced a draft set of guidelines (6).

Clearly the more guidelines and controls are observed at the stage when digital resources are being created, the easier the task of managing and preserving those digital resources becomes and the more attractive the "migration" strategy becomes.

4.1.3 Management Prior To Deposit

The third key factor relates to how that category of digital resource was managed and/or distributed prior to deposit.

As indicated in section (3.5) above, the system used to manage a specific category of digital resources may be very simple or very complex.

At its simplest it could comprise a minimum amount of attribute data held in the header of a file containing a simple text file.

At its most complex it could be a relational or object oriented database management system used to manage a considerable volume of attribute or index data about a very large collection of digital resources. It may have been used to hold thousands of links between data within different digital resources. Finally, it may have been used to provide users with powerful tools to navigate through the contents of the digital resources that it managed.

As a general rule, the simpler and the more standardised the system used to manage the digital resource, the easier and more cost effective it is to implement a “migration” strategy. The more complex and the more proprietary the management system then the more difficult and more costly it is to implement a migration strategy without losing some of the links and other attribute data.

If the management system is open and follows international or de facto standards then – even if the management system is complex and a lot of attribute data is held – provided sufficient time and resources can be made available - a “migration” strategy can usually be devised and implemented.

If the management system is proprietary and no longer supported and it holds a vast amount of complex attribute data and links between data, then “migration” of the management data may not prove cost effective and the “technology preservation” or “technology emulation” approaches may have to be adopted.

4.1.4 Deposit

The fourth key factor relates to how that category of digital resource was actually deposited at the data centre.

Ideally, as indicated in section (3.5) above, the digital resources will be live and the data collection manager will be in contact with the creators prior to deposit. The collection manager will be able to discuss the options with the creator/owner of the data and arrange for the digital resources to be processed and deposited/interchanged by the creators/owners in the most appropriate format. The only limiting factor may be the export options supported on the creator/owner’s system.

This will increasingly be the case in future where the creation process is under the control of the data centre and authors know that they will have to deposit the data in one of a number of approved formats to meet the requirements of the data centre. Both the PRO and Government Departments and the Research Council data centres and individual researchers respectively, should have this kind of relationship in place in future.

At the other extreme, valuable data will continue to be discovered in a range of formats and on a range of media. If they want to hold this data in their collections then data centre managers will have to look at how best to try and preserve such “uncontrolled deposits”.

Clearly, cases where the deposit process is under the control of the data centre should lend themselves more easily to a “migration” strategy. In cases where digital resources are discovered and there is little or no control on the deposit process then “migration” may prove technically difficult and economically non-viable. An example here would be digital resources which have not been documented and which have been written to a specific digital medium using software which prevents the data being copied off that medium. If the data cannot be validated and cannot be copied off that medium then the digital resource is “hardware – dependent” and a “migration” strategy would not be a feasible option. In those cases then “technology preservation” or “technology emulation” may be the only strategies available.

4.1.5 How Well Was the Digital Resource Documented?

The fifth key factor relates to how well the digital resource was documented prior to deposit at the data centre.

As indicated in section (2.2.2) above, adequate documentation is vital to ensure that data centre staff can interpret the data deposited.

All established data centres appreciate the vital importance of documentation. As a result they have prepared guidelines and specifications for all those preparing data and documentation for deposit at their data centres.

Data centres require a minimum amount of documentation to enable them to interpret the data held in the data set, to validate it and to migrate it as required.

If a digital resource is not adequately documented then the data centre may reject it if they have no resources available to document it themselves. Alternatively they may opt for a "technology preservation" or "technology emulation" strategy as a pragmatic short to medium term solution.

If a digital resource is well documented then the data centre will be able to assess which subset of the "migration" strategy is most appropriate for that digital resource.

The Data Archive at the University of Essex (7) publish their "Guide to Depositing Data" on the Web. They are the largest national repository of research data in the social sciences and humanities in the UK. They were set up in 1967. They recommend that data should be copied and deposited as soon as the final version of the dataset has been created. Documentation should be generated as the project proceeds. They ask all depositors to carry out the following tasks:

- Generate copies of the files in machine and software independent format with, where appropriate, data definition files;
- Copy all relevant documentation preferably as machine-readable text files. If no machine readable documentation is available, paper copies clean enough for photocopying may be sent;
- Include a clean original data collection form or an example of the source from which the data was generated;
- Complete a set of Archive deposit forms;
- Send documentation and data by registered post to the Archive or arrange to send machine-readable material by electronic data transfer.

4.1.6 What Were the Results of the Data Validation Exercise?

The sixth key factor relates to the results obtained at the validation stage. All digital resources deposited at a data centre should be validated.

Greenstein (3) defined three validation procedures designed to ensure a digital resource's integrity:

Data assessment (testing the digital resource's completeness; function and consistency)

Data copying (making additional copies of the digital resource to guard against the loss or corruption of any one copy)

Media refreshment (periodically copying one copy of a digital resource onto fresh media to protect against the corruption and content loss which may result from media deterioration)

The first two should be conducted at the deposit stage to ensure the integrity of the digital resource when it was deposited. It is vital to establish that it is complete and that it functions as it was intended to function. If the digital resource is found to be incomplete or malfunctions then the collection manager should go back to the depositor - if that is feasible - and get the problem identified and resolved. If that is not possible then the problems may rule out "migration" as a strategy.

It is also vital to establish that the data can actually be copied to a different medium without corrupting the data or losing the ability to access it. Cases where this may be difficult may include digital publications that are not adequately documented. The digital resource may have been authored and published on a specific digital medium using software that prevents the data being copied off that medium. If the data cannot be copied off that medium then the digital resource is "hardware – dependent" and a "migration" strategy would not be a feasible option. In those cases then "technology preservation" or "technology emulation" may be the only strategies available.

4.1.7 What Usage Restrictions/Requirements Were In Operation?

Depositors of digital resources may impose restrictions on the usage of their data. Some depositors might insist that users can only view or print their data – they should not be allowed to process or alter it in any way. A similar requirement might be that the digital resource must always be presented in the way intended by the author e g certain aspects of the format cannot be altered.

Both these requirements would tend to lead the collection manager to a "migration" strategy built around the "change media" subset or the "convert to standard formats" subset. Digital resources may be converted to PostScript page files or Adobe PDF files or even to page images in TIFF.

There is one case where these requirements could lead the collection manager away from a "migration" strategy. This would be where the digital resource was created on a proprietary application program and held in a proprietary processable format. If the only way the digital resource could be interchanged involved the loss of formatting or presentation data (ASCII; RTF) then this may not prove acceptable. In that case the collection manager might have to adopt a "technology preservation" or "technology emulation" strategy to preserve the presentation or the "look and feel" of the original.

Alternatively, the users of a data centre that manages structured texts may indicate that they need to be able to access the texts in a processable form to edit and annotate them. In this case the collection manager would need to adopt a "migration" strategy built around the "convert to standard formats" strategy and select a standard processable format e g SGML and an agreed DTD.

4.2 Applying the Model

Below the consultancy applies the decision model to the ten categories of digital resources defined for this study. Given the time and resources available for the study a very simplistic approach has been adopted. In reality there would be variations within each category.

The aim of the next ten brief sections is to show data managers how the decision model can be applied. It is not designed to provide definitive guidance. It has always been the view of the consultancy that the people best qualified to make the final decisions on preservation options are the people who understand the specific digital resources and data types to be preserved.

If the model is useful then it should be taken up and applied by the data centre managers. Further work would then need to be done by specialists in each of the ten categories before any detailed guidance could be presented on how best to preserve each category of digital resource.

4.3 Data Sets

The bulk of data sets are taken to comprise a single data type – alphanumeric data. If the data has not already been "worked up" or validated then "working the data up" and "validating" the data can represent a very resource intensive task. If the data is not adequately documented then documenting the data can also represent a major task.

The best approach to preserving data sets is to mandate depositors with documenting them fully and to mandate depositors with depositing them in a platform independent and ideally a software independent format.

The Data Archive at the University of Essex (7) provide the following advice to depositors – "data held in a database management system should be written out as raw data files with identifiable delimiters e g comma delimited files, accompanied by data definition statements where appropriate."

The ideal preservation strategy for data sets is "Digital Information Migration" using the "Convert to standard formats" subset.

4.4 Structured Texts

The bulk of structured texts comprise one dominant data type – alphanumeric data – together with mark-up codes and, increasingly, tagged vector and raster graphics.

Most data centres that specialise in the preservation of structured texts such as literary texts have adopted the "Digital Information Migration" strategy and specifically the "Convert to standard formats" subset.

The Oxford Text Archive follows the Text Encoding Initiative (TEI) guidelines for using SGML to mark up literary works. There are various versions of TEI for different categories of literary work including plays, poems etc. Oxford uses TEI Lite as a basic standard. Clearly there is a significant overhead involved in taking word-processed documents with proprietary mark-ups and tagging them using TEI Lite. However, increasingly, regular depositors are following TEI Lite themselves.

For text only documents a well-equipped centre such as Oxford should not need to rely on any other preservation strategy.

Problematic areas would be where a collection of texts were deposited that comprised text plus vector and raster graphics and the tags or links between the objects were held in a proprietary desktop publishing application or in a proprietary management system. If the desktop publishing application or the management system became obsolete then the centre would need to preserve the operating environment that they ran on until such time as they developed a migration strategy that allowed them to convert the digital resource into a format that could be migrated forwards.

4.5 Office Documents

Office documents cover almost all data types. The bulk would comprise alphanumeric data and raster and vector graphics. Increasingly there may be sound and moving video clips as well but these are covered separately below.

The Public Record Office draft guidelines (6) are designed to address the long-term storage of electronic records created on office systems. They recommend the "Digital Information Migration" strategy and specifically the "Convert to standard formats" subset. The specific formats they recommend are PostScript; Adobe's PDF; TIFF; SGML and Comma Separated Variable ASCII for alphanumeric data.

Given their records management background the PRO favour "formatted forms" where the format and presentation of the document as well as its content is preserved. As a back up they would also endorse the "Change Media" subset where formatted page images of digital documents could be output to paper or microfilm.

For modern digital office documents there should be no need to adopt any "Technology Preservation" or "Technology Emulation" strategies. A pragmatic short term solution for digital resources created on the leading de facto office applications would be to rely on the "Backward Compatibility" of the application. However, this should only be seen as a short-term solution and should always be backed up by converting the documents to a standard formatted form or changing the media.

Two main issues face data centre managers preserving images of paper documents.

The first issue – if the documents are not held in digital form at the time of deposit and hence have to be scanned and digitised – is what resolution should the images be scanned at and should they be captured and held as black and white; greyscale or colour images? The options are listed in section (3.4.1). The answer varies depending on the material to be captured. In addition, the cost of all the options also needs to be taken into account. High-resolution images take a long time to capture and create large file sizes. For standard office documents the "de facto" standard is 200 – 300dpi resolution, black and white or bitonal image capture.

The second issue relates to what file format the image data should be stored in, whether or not the image data should be compressed and, if so, what compression algorithm should be used. The main options are listed in section (3.4.1). For standard office documents the "de facto" standard tends to be the CCITT Group IV facsimile compression algorithm and the compressed data is stored in TIFF.

A problem area identified by the PRO is where - for a given collection of office documents - all the vital attribute data and context data plus the links between the documents were held in a proprietary office suite application or in a proprietary groupware database or document management system. If the office suite, groupware or document management system became obsolete then the centre would need to preserve the operating environment that they ran on until such time as they developed a migration strategy that allowed them to capture the links in a standard form. The effort involved in recreating the links on a preferred open management system would be prohibitive in most cases.

4.6 Design Data

Design data or product data can again cover almost all data types. The bulk would comprise vector graphics (CAD); raster graphics (old manual drawings) and alphanumeric data (text documents plus attribute data in databases. Design data can be held in two and three-dimensional formats.

Data centres managing this category of digital resource recommend the "Digital Information Migration" strategy and tend to rely on a combination of the "Backward Compatibility", "Interoperability" and "Convert to Standard Formats" subsets. For interchange they tend to rely on the de facto "DXF" format or "IGES" for two and three dimensional vector graphics. Standard formatted forms include HPGL; Encapsulated PostScript and TIFF.

Design data managers hold parts catalogues, technical manuals, standards and procedures in SGML and hold images of old drawings and reports etc in TIFF Group 4 format. As a back up many also endorse the "Change Media" subset where formatted page images of digital documents could be output to 35mm microfilm.

For modern design data there should be no need to adopt any "Technology Preservation" or "Technology Emulation" strategies. A pragmatic short - term solution for digital resources created on leading CAD applications would be to rely on the "Backward Compatibility" of the application. However, this should be seen as a short - term solution and should be backed up by converting the documents to a standard formatted form or changing the media.

For older design data then "Convert to Standard Formats" is the preferred subset. "Technology Preservation" or "Technology Emulation" would be relied on only where valuable design data 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.

One problem area would be where - for a given product - all vital attribute and context data plus the links between the design data and documents were held in a proprietary Product Data Management system. Increasingly the trend would be to build the PDM system on top of a standard database engine and to hold all the design data and documents in standard formats.

If a proprietary PDM system was used and became obsolete, the centre would need to preserve the operating environment the PDM system ran on until they developed a migration strategy. The migration strategy would need to allow them to capture the links in a standard form and export the documents and design data in a standard form. The effort involved in recreating the links on a preferred open PDM system would be prohibitive in most cases.

4.7 Presentation Graphics

Presentation graphics can cover a range of data types. The bulk would comprise vector graphics (CAD); raster graphics; animation (moving graphics) and alphanumeric data. A judgement could be taken on how valuable the moving graphics were e g a PowerPoint presentation with build-up graphics could be archived as a series of static frames without much loss of value.

Data centres managing this category of digital resource would be able to adopt the "Digital Information Migration" strategy and rely on a combination of the "Backward Compatibility", "Convert to Standard Formats" and "Change Media" subsets. Standard formatted forms include PostScript; Adobe PDF and TIFF.

As a back up, data centres could adopt the "Change Media" subset where formatted page images of digital documents could be output to microfilm or paper.

For presentation graphics created on de facto standard packages there should be no need to adopt any "Technology Preservation" or "Technology Emulation" strategies. A pragmatic short term solution for digital resources created on the leading de facto presentation graphics applications would be to rely on the "Backward Compatibility" of the application. However, this should be seen as a short - term solution and should be backed up by converting the documents to a standard formatted form or changing the media.

For older presentation graphics then "Convert to Standard Formats" is the preferred subset. "Technology Preservation" or "Technology Emulation" would be relied on only 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.

4.8 Visual Images

The bulk of visual images, self evidently, will comprise one dominant data type – raster graphics data. Where large collections are held then attribute data relating to the images will be held as alphanumeric data and managed in a database.

Most data centres that specialise in the preservation of visual images have adopted the "Digital Information Migration" strategy and specifically the "Convert to Standard Formats" subset.

There are two main issues which face data centre managers preserving visual images.

The first issue – if the images are not held in digital form at the time of deposit and hence have to be scanned and digitised – is what resolution should the images be scanned at and should they be captured and held as black and white; greyscale or colour images. If the decision is colour then what coding scheme should be adopted. The options are listed in section (3.4.1).

A number of excellent articles and text books have been written on this issue (8 – 12). The answer varies depending on the range of material to be captured as images. In addition the cost of all the options also needs to be taken into account. High-resolution colour images take a very long time to capture and create very large file sizes.

The second issue relates to what file format the image data should be stored in and whether or not the image data should be compressed and, if so, what compression algorithm should be used. The main options are listed in section (3.4.1). For visual images a well-equipped centre should not need to rely on any other preservation strategy.

4.9 Speech/Sound Recordings

It is important to distinguish between speech and all other sounds. Speech has a semantic content. 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. Standard digital telephony employs pulse code modulation and logarithmic coding and a sampling rate of 8 kHz and 8 bit code words is used to give a bit rate of 64 kbps. The bandwidth needed to transmit or record speech is smaller than that needed for other sounds e g music due to the narrower range of frequencies used when talking. Hence the number of bits needed to code 1 minute of speech is less than the number of bits needed to record 1 minute of music.

- Speech synthesis – the translation by computers of a coded description of a message into speech e g computers talking to people.

- Speech recognition and speech understanding. The individual components of speech – the words – are recognised. This facilitates people talking to computers and dictating text or issuing commands etc.

Non speech sounds do not generally have a complex semantic context. Non speech sound processing can therefore be divided up into two areas:

- Sound coding – 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. Waveform sound can be uncompressed as on audio compact discs or compressed. Any sound, including speech and music, can be recorded in this way. Compact Disc Digital Audio recording employs pulse code modulation and linear coding and a sampling rate of 44.1 kHz and 16 bit code words. The CD Audio standard supports stereophonic sound over two channels. Hence the aggregated bit rate for a stereophonic CD audio bitstream is 1.411 Mbps.

- 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. MIDI is a much more compact coding technique than digitised samples.

The sound files which data centre managers need to preserve will generally contain sound data coded either as a digitised analog sound signal or as notes for a MIDI instrument. A MIDI file usually has a MID file extension. A WAVE or WAV file is the “de facto”

standard for storing an analog signal in digital form under MS Windows.

This is a very specialised area where individual studies need to be conducted by experts in the field of digital speech and sound recording and processing.

It appears that data centres managing this category of digital resource would be able to adopt the "Digital Information Migration" strategy and rely on a combination of the "Backward Compatibility", "Convert to Standard Formats" and Change Media subsets.

For sound files created on de facto standard packages there should be no need to adopt any "Technology Preservation" or "Technology Emulation" strategies. A pragmatic short - term solution for digital resources created on the leading de facto applications would be to rely on the "Backward Compatibility" of the application. However, this should be seen as a short - term solution and should be backed up by converting the documents to a standard form.

4.10 Video Recordings

The bulk of digital video recordings, self evidently, will comprise one dominant data type – motion video or moving image data. Increasingly digital video resources also contain interleaved audio data as well. Where large collections are held then attribute data relating to the moving images or interleaved audio and video data will be held as alphanumeric data and managed in a database.

This is a very specialised area and is still a relatively new area where individual studies need to be conducted by experts in the field of digital video recording and processing.

Most data centres that specialise in the preservation of film and video recordings are still in the early stages when it comes to collecting and preserving digital video data. The bulk of holdings will still be in analogue film or videotape formats. Where practical, such centres have adopted the "Digital Information Migration" strategy and specifically the "Convert to Standard Formats" subset. The MPEG standards provide several standards for the compression of full motion video.

However, this is an area where early digital video material was inevitably created on proprietary applications and held in proprietary formats. Where the applications are now obsolete and unsupported then centres would have to turn to a "Technology Preservation" or "Technology Emulation" strategy to preserve the data until such time as they can work out an acceptable migration strategy.

4.11 Geographic/Mapping Databases

Geographic/mapping data can cover almost all data types. The bulk would comprise raster graphics (base mapping data) and vector graphics and alphanumeric data (attribute data – links etc in databases). Geographic data can be held in two and three-dimensional formats.

This is a very specialised area where individual studies need to be conducted by experts in the field of GIS systems and mapping data.

Chris Perkins has already helpfully arranged cartographic software packages in increasing order of difficulty and functionality (13) ranging from atlases and route planners right up to full Geographical Information Systems (GIS).

Data centres managing this category of digital resource recommend the “Digital Information Migration” strategy and rely on a combination of the “Backward Compatibility”, “Interoperability” and “Convert to Standard Formats” subsets with “Change Media” as the back up option.

For modern mapping data there should be no need to adopt any “Technology Preservation” or “Technology Emulation” strategies. A pragmatic short - term solution for digital resources created on the leading de facto mapping applications and GIS systems would be to rely on the “Backward Compatibility” of the application. However, this should be seen as a short - term solution and should be backed up by converting the data to as standard a form as possible.

“Technology Preservation” or “Technology Emulation” would need to be relied on where valuable data 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.

The main problem area is where - for a given GIS database - all the vital attribute data and co-ordinates and all the links between the vector layers and the raster data are held in a proprietary GIS system which becomes obsolete. Here the centre would need to preserve the operating environment that the GIS system ran on until they developed a migration strategy. The migration strategy would need to allow them to capture the co-ordinates and the links in a standard form and export the documents and data in a standard form. The effort involved in recreating the links on a preferred GIS database would be very high in most cases.

4.12 Interactive Multimedia Publications

By definition multimedia publications comprise at least three data types. Most multimedia publications will comprise motion video and audio data interleaved; many will comprise animation and interleaved audio data and most will comprise some still images and graphics and alphanumeric data. Most of the early multimedia publications will have been produced on one of the CD formats and will have been authored and edited using proprietary multimedia editing and authoring packages. They will be accessed via proprietary access software.

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

It is a fact that many valuable multimedia publications were created using proprietary editing, authoring and access software which is now no longer supported and which does not work on current hardware and software platforms. Given that fact, then data centre managers will have to adopt a “Technology Preservation” or “Technology Emulation” strategy to preserve the data until they can develop a practical migration strategy. It may prove impossible to migrate the data in future without the loss of considerable data.

4.13 Summary

The following table summarises the results gained in this first attempt to apply the model to the ten categories of digital resources defined for this study.

	Digital Resource	Preservation Strategy	Subset Of Strategy	Notes
1	Data Sets	Digital Information Migration	Convert To Standard Formats;	
2	Structured Texts	Digital Information Migration	Convert To Standard Formats;	
3	Office Documents	Digital Information Migration	Convert To Standard Formats; Backward Compatibility; Change Media	
4	Design Data	Digital Information Migration	Backward Compatibility; Interoperability; Convert To Standard Formats; Change Media	Technology Preservation/ Emulation as short term strategy for product data on obsolete systems;
5	Presentation Graphics	Digital Information Migration	Backward Compatibility; Convert To Standard Formats; Change Media	
6	Visual Images	Digital Information Migration	Backward Compatibility; Convert To Standard Formats; Change Media	
7	Speech/Sound Recordings	Digital Information Migration	Backward Compatibility; Convert 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; Convert 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; Convert 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;

5 Developing a Cost Model

5.1 Identifying the Cost Elements

5.1.1 Issues & Preferred Methodology

In chapter four the consultancy has drawn up a decision model that can be used by data centre managers when assessing the agreed ten categories of digital resources to determine the most appropriate method of long term preservation for each category. The consultancy identified seven key factors, which need to be taken into account when deciding on the preferred preservation strategy for each category of digital resources. The model was applied and the consultancy drew up a table listing the preferred preservation strategies for each of the ten categories of digital resources covered by this study.

The terms of reference for this study also called on the consultancy to build on the decision model and go on to draw up a cost model that can be used to establish and compare the costs of the preferred methods of preservation for each category of digital resource.

In chapters two and four above, the consultancy described Greenstein's framework (3) which defines seven modules, which must be addressed by any digital collection policy, of which preservation is one. The seven modules and key sub modules were listed in a table in section (4.1.1) above.

To draw up a scientifically valid cost model the consultancy would have had to carry out two major tasks. The first would have been to analyse the seven modules in the framework and identify (for a range of data centres/archives/digital libraries all managing the same category of digital resources) all the cost elements contained within each module. The total set of costs obtained from each data centre/archive/library would represent the bulk of the costs involved in managing a specific category of digital resource in that data centre/archive library. An average figure would then need to be calculated provided sufficient standardisation of procedures had been discovered to make this a meaningful figure.

The second major task would then have been to take this average total cost and decide what percentage related to "preservation" as opposed to all the other functions and roles of a data centre, data archive or digital library. Does the cost of the creation of an online catalogue or a printed catalogue constitute a "preservation" cost?

For the results of the first task to be scientifically valid a much longer and better resourced study would need to be commissioned involving experts in all ten of the categories of digital resources covered in this study. The extended study would have to cover at least 10 – 20 centres, which were all preserving the same category of digital resources and following similar procedures. This would represent a total of up to 200 centres. This current short study did not allow the consultancy time to visit one data centre for each category of digital resource.

The results of the second task cannot be scientifically valid or meaningful until some overall consensus is reached on how the costs of preservation can be separated from the costs associated with the other six main modules in Greenstein's Framework.

Greenstein himself (3) points out that this is very difficult to do at this early stage in the development of data archiving, given the way in which all the modules of the framework

appear to overlap with and depend on each other.

Given these difficulties and the time and resource constraints of the study, the consultancy has adopted the following pragmatic methodology:

Firstly, based on interviews and desk research, the consultancy analyses the seven modules in the framework and tries to identify all the generic cost elements contained within each module, which relate directly or indirectly to the preservation of digital resources. The total set of costs identified represent the bulk of the costs involved in preserving one or more categories of digital resource in data centres/archives/digital libraries.

This provides a subjective and simplistic initial cost model as shown in graphic form in section (5.1.9) below. Hopefully this will serve as the basis for discussion and debate from which a greater consensus will emerge on how best to attribute costs.

Secondly, the consultancy then attempts to repeat the exercise in more detail for four specific categories of digital resources out of the ten categories of digital resource covered in this study. Further studies will, hopefully, build on this to provide more scientific cost models for each agreed category of digital resource which data centre managers can then begin to use with some confidence when drawing up budgets and project plans.

5.1.2 Creation Costs

5.1.2.1 Level of Control

The first question to ask about the creation process is can the data centre/archive/library exert any influence over how the data that it manages is created? Depending on the type of data centre/ data archive/ digital library they are managing and how it is funded, collection managers may have no influence over how the data is originally created or they may have considerable control over the creation process. Some examples can best illustrate this point.

A Managing Funded Research Data

The Natural Environment Research Council (NERC) is one of several major research councils in the UK. One of their major roles is to "promote and support high quality, basic, strategic and applied research" (14). They, therefore, fund a considerable volume of research projects. NERC recognise that environmental research involves the collection of data and that the subsequent management of the data is a vital part of NERC's mission. NERC has therefore established some seven designated Data Centres and the NERC Data Strategy Group has published a NERC Data Policy Handbook.

The 7 data centres have gained considerable expertise in the management of environmental data and are in a position to define and recommend good practice in the areas of data gathering, data management and data preservation. Given that NERC fund the original research projects and also fund the data centres there is clearly scope for advising individual research projects on how best to gather and manage their data and how to deposit it to the relevant centre. In some cases minimum standards of data management can be imposed.

There are many similar examples where the lifecycle of the research data can be controlled in a loop. As more experience and expertise is gained in how best to manage the various

categories of research data so there will be more scope for imposing minimum standards of data management on funded research projects.

B Managing Government Records

The Public Record Office is the official repository for government records. As a growing volume of government records are created and managed in digital format so the PRO will be called upon to select them and manage and preserve them in digital format. For the PRO to provide this service in future there is clearly a need to agree standard procedures and standard interchange formats with government departments. Hence the PRO have issued the draft EROS guidelines (6), which are being considered by Departmental Records Officers and other interested third parties.

This is another example of a controlled lifecycle where the archive will be able to influence how the records are created and where a continuous dialogue will be developed between depositors and the collection managers.

C Managing Scholarly & Academic Resources

The Arts and Humanities Data Service (AHDS) is a recently established national service funded by the Joint Information Systems Committee (JISC) of the UK's Higher Education Funding Councils to "collect, describe and preserve the electronic resources which result from research and teaching in the humanities"(3). To date AHDS comprises a managing executive and some six data centres managing digital resources covering the areas of archaeology; history; literary, linguistic and other textual studies, the visual arts and the performing arts.

One of the key objectives of AHDS is to "promote awareness amongst the scholarly community about the importance and value of electronic information and provide guidance in its effective creation and use".

Another objective of AHDS is to "facilitate fruitful partnerships between scholarly communities and the commercial and not for profit information services and funding agencies upon which they increasingly rely in order to enhance the production and preservation of high quality digital resources and to provide more uniform access to them".

Clearly AHDS will not fund all scholarly research in the humanities and hence the links between the data centres and their potential depositors will not be quite as tight as those between NERC and its funded research projects.

AHDS serve a specific community and hence they have already been able to target their community effectively and explain to them the benefits of depositing their digital resources with AHDS centres. Increasingly the centres will also be able to advise scholars and researchers in the Humanities on the best practices for creating and documenting and depositing their digital resources.

D Managing Uncontrolled Deposits & Discovered Resources

Outside of these close communities, where the data centres can exert a strong influence over the creators of the digital resources, many other digital archives and libraries have no influence on their depositors.

There will continue to be many cases where valuable digital resources will be discovered or deposited and will not have been adequately managed or documented. Collection managers in these cases will have to make difficult decisions about whether they can afford to validate,

document and, in many cases, work up these digital resources to the stage where they can be made available as a self contained resource.

5.1.2.2 Overall Cost of Creating Digital Resources

In the first three cases – where the data centre/archive does have some contact with the creators of the digital resource – it is very useful to establish the overall cost of the project.

Before deciding how best to manage and preserve the digital resource created by a project it is useful to establish whether it was a multinational research project that cost several million pounds or a small piece of desk research that cost a few thousand pounds.

This is not to equate the cost of research with the quality of the output but rather to help the data centre manager to justify the expenditure of resources on subsequently managing and preserving the data.

There is scope here for the Research Councils and their data centres to carry out more research on the overall costs of research and the average costs of data collection and management as a percentage of the total cost of the research project. This could lead to the issue of guidelines on what percentage of a project's resources should be devoted to data management.

Within government an efficiency review has been conducted which looked at the costs of:

- creating and managing records while they are active;
- reviewing them and storing the selected records between first and second review;
- reviewing them again and storing the selected records prior to selection by PRO;
- selection by PRO and subsequent accessioning and long term preservation.

Such a review is a necessary precursor to a study, which looked at the costs and benefits of moving to electronic records management and then compared them with the existing costs.

In the academic community, a similar study would look at the true costs of supporting academic research. Included in those costs should be:

- the cost of educating scholars to the point when they can produce such works;
- the costs involved in trying to access previous related research;
- the cost of cataloguing and storing in physical form the results of that research in individual college and university libraries and in centralised collections where they exist;

In all cases it will be discovered that the costs currently devoted to managing the digital resources created by funded research; government departments and scholars is a very small percentage of the overall costs which go in to creating the resources. This should help make a strong business case for increasing the funds available to data centres/archives in future.

5.1.2.3 Best Practice at Creation Cuts Management & Preservation Costs

The most important area to look at is how the adoption of "best practice" at the data gathering

and creation stage can help simplify the task of managing and preserving the digital resource in future. Simplification will result in reduced costs.

One consistent message came out of all the interviews conducted by the consultancy with data centre managers. The biggest cost which they all face is trying to "clean-up" or "work-up" digital resources, which should have been cleaned up or worked up at the time they were created. The creators of a digital resource are best equipped to validate it and to document it. If they do not do this then the cost of "clean – up" at a later stage when most of the context will have been lost is conservatively estimated to be ten times greater. In many cases it is impractical to attempt to clean - up digital resources retrospectively and the digital resources would be rejected.

The NERC data centres make the point very strongly that best practice for the collection, processing, validation, management and documentation of scientific data sets gathered as the result of scientific experiments, surveys etc is absolutely vital to the efficient management of those data sets in future.

For the purposes of this study then "creation costs relating to preservation" covers all costs incurred in collecting, processing, validating, managing and documenting digital resources up to and including the deposit of the digital resource with the data centre. The creator/owner of the digital resource would normally be liable for those costs.

If one body funds the research and the data centre, then it should, in theory, be easier to argue the case to that funding body that money spent at the creation stage can pay dividends at the management stage. A strong business case can be made that if £A is spent on managing the data professionally during the creation process then £A x 10 will be saved on managing and preserving the data at the data centre in future.

The general figure that is used in the research council environment is that between 2 – 5% of the cost of the research project should be devoted to data management. In the majority of cases a figure of 3% was quoted to the consultancy. On major scientific projects the data centre may actually assign a member of staff to that project to gather and process and document the data during the project.

In the case of government it is interesting to note that the PRO are beginning to adopt a similar approach. The Government Services Division (GSD) staff will be working much more closely with Departmental Records Officers and IT staff to agree procedures and guidelines for the management of electronic records by Departments and the creation of finding aids etc by Departments prior to deposit of the records with PRO.

Already the AHDS data centres are devoting considerable resources to the promotion of good practice for depositors. They are issuing guidelines for depositors and a series of best practice guides advising scholars and researchers on the preferred application programs to use, the preferred formats to store their digital resources in and the level of documentation required. The more success stories emerge and the more support is given to AHDS by funding bodies in the academic community then the more the scholarly community will look to AHDS for advice on how best to manage digital resources prior to deposit.

5.1.2.4 Data Centre Costs Relating to Creation Practices

So far in this section the consultancy has made the following points:

- different types of data centre/archive/library have more or less control over how the data they manage is created;
- it is useful to establish how much money is spent creating the digital resources deposited at a specific data centre/archive/library. It can help justify increasing the budget for the data centre and for agreeing a percentage figure to be spent on managing the data efficiently;
- good practice at the creation stage can save considerable resources at the data management and preservation stage. The accepted figure is that it costs ten times as much to correct bad practice retrospectively than it would to have adopted good practices at the creation stage.

Given these points, data centre managers cannot ignore the creation stage – even if it is outside their control. Data centres/archives/libraries face two costs relating to the creation stage.

The first cost they should budget for is the cost of promoting good practice to their depositors. This can include posting guidance notes on their Web pages; running courses for defined user groups; educating funding bodies and including guidance notes in funding literature. The costs will vary depending on the type of data centre/archive/library and the clientele they serve. This can be seen as a preventative measure and all data centres etc should invest some resources in this area.

The second cost they should budget for is the cost of correcting mistakes and examples of bad practice at the creation stage. Arguably the less they spend on the first cost the more they will have to spend on the second cost.

Clearly, on a per digital resource basis, the second costs will far outweigh the first costs. Hence again it is enlightened practice to invest in education to try and reduce the number of instances of bad practice.

There is theoretically no limit to the amount of money that data centres etc could spend on "cleaning-up" deposited digital resources. There are three practical ways of limiting the costs.

Firstly if the data centre has a fixed budget it must allocate a maximum figure to "clean-up" activities.

Secondly it must define a basic standard. Valuable resources that fall below that standard would be brought up to that standard but would not be taken beyond that standard.

The third approach would be to state that digital resources, which do not meet a minimum standard in areas such as documentation, would not be accepted.

Both the promotional (best practice) and the corrective (clean-up) costs relate directly and indirectly to preservation. Examples of the resources devoted to the promotion of best practice and to cleaning up poor data by data centres are provided below when looking at the preservation costs for specific categories of digital resources.

5.1.3 Selection & Evaluation (Acquisition) Costs

The actual selection of digital resources will be based on the centre's collection policy and an assessment of the content and quality of the specific resource. The costs associated with this exercise do not relate directly or indirectly to preservation.

The evaluation of digital resources does involve assessing them against a series of technical and practical criteria. How easily can these resources be managed, catalogued, accessed by end users and preserved by the data centre?

The costs associated with evaluating a digital resource do therefore relate directly or at least indirectly to preservation.

The cost of evaluating a digital resource will depend on the size and complexity of the digital resource and how well documented it is. Examples are provided below when looking at the preservation costs for specific categories of digital resources.

5.1.4 Data Management Costs

Data management covers all the tasks involved in managing a digital resource once it has been accepted into a collection. Greenstein breaks data management down into – data structure; documentation; storage and validation. In reviewing the costs the consultancy uses these four subdivisions but addresses them in a different order to Greenstein.

5.1.4.1 Documentation

The first task involves checking the documentation supplied with the digital resource, editing it or adding to it, if required, and then managing the documentation. Increasingly, because centres are providing users with online access to the digital resources, there is a need to hold the documentation in digital format as well. Hence if the documentation is only provided on paper there is an extra task and cost involved in digitising it.

Again, the more resources the centre puts into promoting good practice to depositors the better the documentation should be and hence the lower the costs associated with reading, amending and managing the documentation.

The documentation should describe the resource's structure, contents, provenance and history. Most of the NERC and AHDS data centres hold documents on their Web sites which provide guidance to depositors on what documentation they need to provide. Some provide documentation templates for depositors to complete (7).

Even when a complete set of documentation is provided, the resources required to study the documentation is significant. When the documentation is poor then clearly the costs increase dramatically as the centre has to test the digital resource and produce additional documentation. Many of the centres visited by the consultancy did not have the resources to allow them to do this. If a digital resource was not adequately documented it would be rejected.

Managing the documentation and converting it into digital format can prove very costly, particularly if the data centre hold a large number of digital resources and there is a significant volume of paper documentation for each resource.

The costs associated with reading, editing and managing the documentation for a digital resource do relate directly or at least indirectly to preservation. Without documentation it is impossible to know whether a resource can be preserved and to determine the preferred preservation strategy for that resource.

Examples of the resources devoted by data centres to reading, editing and managing documentation are provided below when looking at the preservation costs for specific categories of digital resources.

5.1.4.2 Validation

The second task covers a number of procedures, which are together designed to ensure the integrity of the digital resource.

The key is assessment. After the documentation has been studied the resource has to be assessed to ensure the following:

- that the resource is complete as documented;
- that the resource is functioning properly and operates on the specified hardware and software environments;
- that the resource is consistent;

The resources available to the centre will determine how exhaustive the validation process is. If the resource comprises one data type and it can be processed in a standard way then the testing need be less exhaustive than with a complex resource.

Greenstein uses validation to also cover the task of copying the digital resource to several copies of the same medium or to several different media to provide a back-up or contingency in the event of one piece of media being destroyed or damaged. He also uses validation to cover the task of periodically refreshing the media – moving the copies of the digital resource to newer, fresher storage media to protect against the corruption that would result from any deterioration of the media.

There is a lot of iteration involved and hence costs are repeated several times during the life of a resource. Refreshing may be done on a yearly basis for all copies of the digital resource held. After refreshing then the digital resource would be assessed again to ensure no corruption took place at the refreshing stage.

The costs associated with assessing, copying and refreshing a digital resource all relate directly or at least indirectly to preservation. Without assessment it is impossible to know whether a resource can be preserved and to determine the preferred preservation strategy for that resource.

Examples of the resources devoted by data centres to assessing, copying and refreshing digital resources are provided below when looking at the preservation costs for specific categories of digital resources.

5.1.4.3 Data Structure

The third task covers how a digital resource is formatted, compressed and encoded and how it had been managed when it was deposited. It also then covers any changes which the data centre decided to make to the compression, encoding and format prior to storing and

managing the digital resource in the centre.

The following factors will impact on the amount of resources required for structuring the digital resource:

- How the digital resource was created;
- How the digital resource will be accessed and used by the clientele;
- The data centre's preferred digital information management strategy

If the digital resource was created on a de facto application and the data centre supports a "Backward Compatibility" subset of the "Digital Information Migration" strategy then the minimum of costs will be needed in the short to medium term for data structuring.

If the digital resource was deposited in one of the standard processable formats supported by the data centre then again the minimum of costs will be needed in the short to medium term for data structuring.

If the digital resource was created on a proprietary application and deposited in a proprietary format not supported by the data centre then significant costs would be involved in converting the digital resource into a standard format for long term storage and management.

The costs associated with converting a digital resource into a preferred format for management and long term preservation do all relate directly or at least indirectly to preservation. The costs associated with converting a digital resource into a preferred format for delivery to users do not relate directly or indirectly to preservation.

Examples of the resources devoted by data centres to converting the structure of digital resources are provided below when looking at the preservation costs for specific categories of digital resources.

5.1.4.4 Data Storage

The fourth task covers the computer hardware and the storage media used to store a digital resource once the resource has been accepted into a collection. The storage options are impacted by the resources available, by the volume of data to be stored and by how it will be used and preserved.

Greenstein identifies four generic data storage scenarios:

- Data warehoused off-line on behalf of some third party and only delivered to that third party in the event of their experiencing some unrecoverable corruption or data loss.

- Data stored off-line or near-line and only distributed to users upon request either via pre-arranged network transmission procedures or on some hand held digital distribution medium (CD ROM; diskette; compact tape format etc);

- Data stored on-line and distributed (via anonymous file transfer or the world wide web) or browsed/analysed (via a Telnet connection or the world-wide web) in real time over a network;

- Mixed distribution scenarios involving some combination of those described above.

The consultancy would emphasise the essential distinction between online or near-line storage for the active management of digital resources and offline storage by the centre at a different location or by a third party for the centre at a separate location.

The on-line or near-line storage would be set up to meet the access requirements of the users and the management requirements of the centre. It may prove more convenient to manage the digital resources online on random access media so they can be re-organised and copied more efficiently. The online system could be managed by the centre or managed for the centre by a third party to a Service Level Agreement (SLA). None of the costs associated with such active management of digital resources relate directly to preservation. A percentage of the costs would be indirectly related to preservation (migration, refreshing etc).

The off-line storage would be set up specifically to meet the preservation requirements of the centre. The off-line storage facility should be managed in a different location to the on-line storage, which meets the requirement for a controlled storage environment and adequate security. The off-line storage facility could be provided by the centre or, more likely, by a third party specialist archive centre to a Service Level Agreement. (SLA).

The costs associated with such an off-line archive storage facility all relate directly to preservation.

Examples of the resources devoted by data centres to off-line archive storage of digital resources and of the indicative prices charged by third parties for such a service are provided below when looking at the preservation costs for specific categories of digital resources and in section (5.6).

5.1.5 Resource Disclosure Costs

Resource disclosure covers how information about a specific digital resource is made available to end-users. What information is available will depend, as we have seen above, on how the resource is documented. However, the costs of documentation have been covered under Data Management above so they are not repeated here.

Greenstein makes the point that how the information is made available to users will depend on the resource discovery tools that are implemented for the collection as whole. Online tools include resource discovery agents, logically ordered gateways and on-line catalogues. An excellent Library & Information Briefing (15) on metadata provides a more detailed review of the options and issues available.

The costs associated with developing and making available online such resource discovery tools do not relate directly or indirectly to preservation.

5.1.6 Data Use Costs

Data use covers the costs associated with delivering the digital resources to end users. The

costs depend upon the structure of the specific digital resource, how they are stored and how the user needs to access them - which may involve the resources being converted into different delivery formats etc.

Greenstein helpfully distinguishes three generic user scenarios again:

- Resources are accessed on-line using client/server technologies and the collection managers manage the server;
- Resources are accessed on-line using client/server technologies and the collection managers do not manage the server (e g resources which are included in a collection but are served to users by a third party under the terms of service level agreement);
- Resources and appropriate software are both resident on workstation to which the user has direct access (e g a plug – and – play CD ROM product)

Greenstein points out that in his opinion collection managers should be able to support two or more of the above options.

The costs associated with delivering the digital resources to end users online or via CD ROM publishing etc do not relate directly or indirectly to preservation.

5.1.7 Data Preservation Costs

Digital preservation has been defined in section (2.2) above. The consultancy divided up digital preservation into three specific tasks:

1 Preserving bit streams through copying/refreshing;

2 Ensuring we can interpret the data by preserving the documentation;

3 Ensuring we can continue to decode the data in future by adopting one of the following three preservations strategies:

- Technology preservation
- Technology emulation
- Digital information migration

The consultancy is following the Greenstein framework in developing the cost model in this section. Hence the costs associated with task one – copying/refreshing – are covered in subsection (5.1.4.2) under Validation and the costs associated with task two – documentation – are covered in subsection (5.1.4.1) under Documentation.

The costs associated with task three are the specific preservation costs which need to be identified here and added to the costs identified in the section above and below as relating to preservation.
The costs associated with task three will clearly vary depending on which preservation strategy is adopted. The table in section (4.13) above indicated that the preferred strategy for 9 out of the 10 categories of digital resources covered by this study was “Digital Information Migration” with varying subsets preferred for specific categories. Hence in this section and

below when looking at the costs of preservation for each of four specific categories, the consultancy concentrates on identifying the costs associated with a “Digital Information Migration” strategy.

5.1.7.1 Technology Preservation

Apart from copying and refreshing, which are covered above, this strategy involves the following tasks:

- preserving the original application program used to create or access the digital resource;
- preserving the operating system software that the original application programs run on;
- preserving the computer hardware platform that the operating system software was designed to run on.

In section (2.2) the consultancy concluded that this strategy cannot be regarded as viable for anything other than the short to medium term. In section (4) the consultancy only recommended the use of “technology preservation” as a relatively desperate measure in cases where valuable digital resources cannot be converted into 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.

This strategy should only be adopted where the only practical way to access a valuable digital resource was via an application that would only run on operating system software that would itself only run on an obsolete hardware platform. In this situation then collection managers would be best advised to seek out a specialist third party (if one could be found) with that hardware environment. They should then run the software and attempt to migrate the data off to at least a software-dependent format and ideally to a software independent format.

Hence the costs involved in adopting this strategy would be of two types.

Firstly, in the short term, when a data centre changed their hardware and software environment, they would incur short term costs. They would incur the cost of keeping the old hardware and software environment running for a short period of time while they worked on a migration strategy to cover those valuable digital resources that could only be accessed via applications that would only run on the old environment.

Secondly, after they have changed their hardware and software environment and switched off the old one - if they discovered any valuable old digital resources that could only be accessed via applications that would only run on the old environment they would pay the cost of using a third party. They would need to identify a third party still running the old hardware and software and pay them to load the application and the data and to convert the data into a standard format which they would be able to preserve and migrate forward on their new environment.

All the costs identified above would relate directly to preservation.

5.1.7.2 Technology Emulation

This strategy has a lot in common with the technology preservation strategy described above.

Apart from copying and refreshing, which are covered above, it involves the following tasks:

- preserving the original application program used to create or access the digital resource.

In addition, as indicted in section (2.2) above, this strategy involves software engineers performing the following tasks:

- designing and running 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.

In section (2.2) above the consultancy concluded that this should be seen as a short to medium term strategy or a specialist strategy where the need to maintain the look and feel of the original digital resource is of great importance to the collection's user base. This strategy could be adopted where the only way to access a valuable digital resource was via an application that would only run on operating system software that would itself only run on an obsolete hardware platform.

In this situation then collection managers would be best advised to seek out a specialist third party (if one could be found) able to emulate that hardware and operating system software environment. They should then run the software and attempt to migrate the data off to at least a software-dependent format and ideally to a software independent format.

Hence the costs involved in adopting this strategy would take the form of a fee to a third party. This would cover the one time use of their facilities to emulate the required hardware and software environment so the centre could run the application and attempt to convert the digital resource to a preferred standard format. They would then preserve the digital resource in that standard format on their current environment and migrate the data forward to future environments.

The costs identified above would relate directly to preservation.

5.1.7.3 Digital Information Migration

Digital information migration is not a simple agreed strategy. It shows the most promise for the future and is the most widely adopted strategy in the centres visited by the consultancy. In section (2.2) the consultancy divided "Digital information migration" into four subsets:

- Change media
- Backward compatibility
- Interoperability
- Convert to standard formats

Each of these subsets can be divided up into a series of tasks, each of which will have a cost associated with it. The full set of tasks for each subset are reviewed below. In a table in section (4.13) above the consultancy has listed which subsets would be most appropriate for each of the individual categories of digital resource covered in this study. These lists are used below when reviewing the costs associated with preserving five specific categories of digital resource.

A Change Media

One basic subset of the migration strategy involves the transfer of digital resources from less stable to more stable media. The most prevalent version of this strategy involves printing digital information onto paper or recording it on microfilm. Paper and microfilm are more stable than most digital media and no special hardware or software are needed to retrieve information from them.

The costs associated with this strategy are threefold:

- Costs associated with formatting the digital resource and either printing it as a series of page images on paper or recording it as a series of page images on roll microfilm or microfiche. For the print option then, depending on the volumes, this could be done in-house or the service could be purchased from a print bureau. For the Computer Output Microfilm (COM) option the service would need to be purchased from a COM bureau. Two copies of the microfilm would be needed – one for active use and one for archive.

- Costs associated with managing the resulting microfilm and paper. These would include indexing and active and archival storage and the provision of microfilm reading and printing equipment.

- Costs associated with making copies of the paper or prints from the microfilm.

Of the costs identified above, the costs relating to the printing or COM recording and the costs related to the archival storage of the paper or microfilm would relate directly to preservation. The costs of storing the active copy of the paper or microfilm and the costs of making copies for users would not be related to preservation.

B Backwards Compatibility

This second subset of the migration strategy relies on popular application software being "backward compatible". The latest versions of most popular application packages will be capable of decoding files created on earlier versions of the same package – particularly the previous two or three versions. The most prevalent version of this strategy simply involves testing the process and then loading files created on previous versions of the application program into the new version and saving them in the new file format.

The costs associated with this strategy (ignoring copying and refreshing costs covered above) would be relatively low in the short to medium term provided the applications remain backward compatible. If the collection manager decided to upgrade all files created on version X of the application to version X + 1 then provided it was feasible to write a macro that retrieved each file and saved a copy of each file in the new format then this process could be automated. The costs would include the following:

- The cost of setting up the macro and running it;
- the cost of checking a subset of the resulting files to ensure no corruption had taken place;
- the cost of deleting the previous version of the files if the collection manager opted to do that.

All the costs identified above would relate directly to preservation.

C Interoperability

The third subset of the migration strategy relies on "interoperability" between rival popular

application programs. Digital resources created on a non preferred/obsolete application program can be exported in a common interchange format and then imported into a preferred current application program that runs on the current hardware and software environment.

The simpler the digital resource the easier it is to interchange the resource between application programs without any significant loss of data and hence the lower the costs incurred. The more complex the digital resource the more difficult it is to interchange the resource between two application programs without any significant loss of data.

If the collection manager decided to export all files created on a non preferred application in a standard interchange format and to then import all or a subset of those files into a preferred application and check them, then this process could be automated.

The costs would include the following:

- the cost of testing the interchange on a range of representative documents;
- the cost of setting up the program and running it;
- the cost of checking a subset of the resulting files to ensure no corruption had taken place;
- the cost of deleting the previous version of the files if the collection manager opted to do that.

All the costs identified above would relate directly to preservation.

D Conversion to Standard Formats

The fourth and most popular subset of the migration strategy is designed to reduce a large number of different formats down to a very small number of standard formats that can still encode the structure and form of the original.

The simpler the digital resource, the easier and hence the cheaper it is to select a standard format and convert the digital resource. In many cases the process would be identical to that described above under “interoperability”. Digital images can be converted from one compression algorithm to another and from one file format to another automatically after initial tests have been completed. Most documents created on Windows applications can be converted to PostScript files or to Adobe’s PDF format.

Converting a text document from proprietary mark-up to TEI Lite would require manual intervention. The exact resources would depend on the length and complexity of the text document.

The main costs associated with this subset would include some of the following:

- Agreeing on the preferred standard formats;
- Testing the conversion for a specific category of resource
- Running the conversion as a batch process
- Testing a sample of converted resources;
- Deleting the old versions if required
- Copying the resulting files

All the costs identified above would relate directly to preservation.

5.1.8 Rights Management Costs

Rights management covers all the processes involved in defining and upholding the rights of the Depositors and the rights of the users of the data centre. Greenstein defines rights as covering intellectual and property rights and related legal issues including data protection and confidentiality.

He points out that the rights vested in a resource may not only determine how the resource can be accessed and used but can also determine how and whether they can legally be preserved by a third party.

For the purposes of this study the costs associated with rights management, which can be substantial, are deemed not to relate directly or indirectly to preservation.

5.1.9 Initial Cost Model

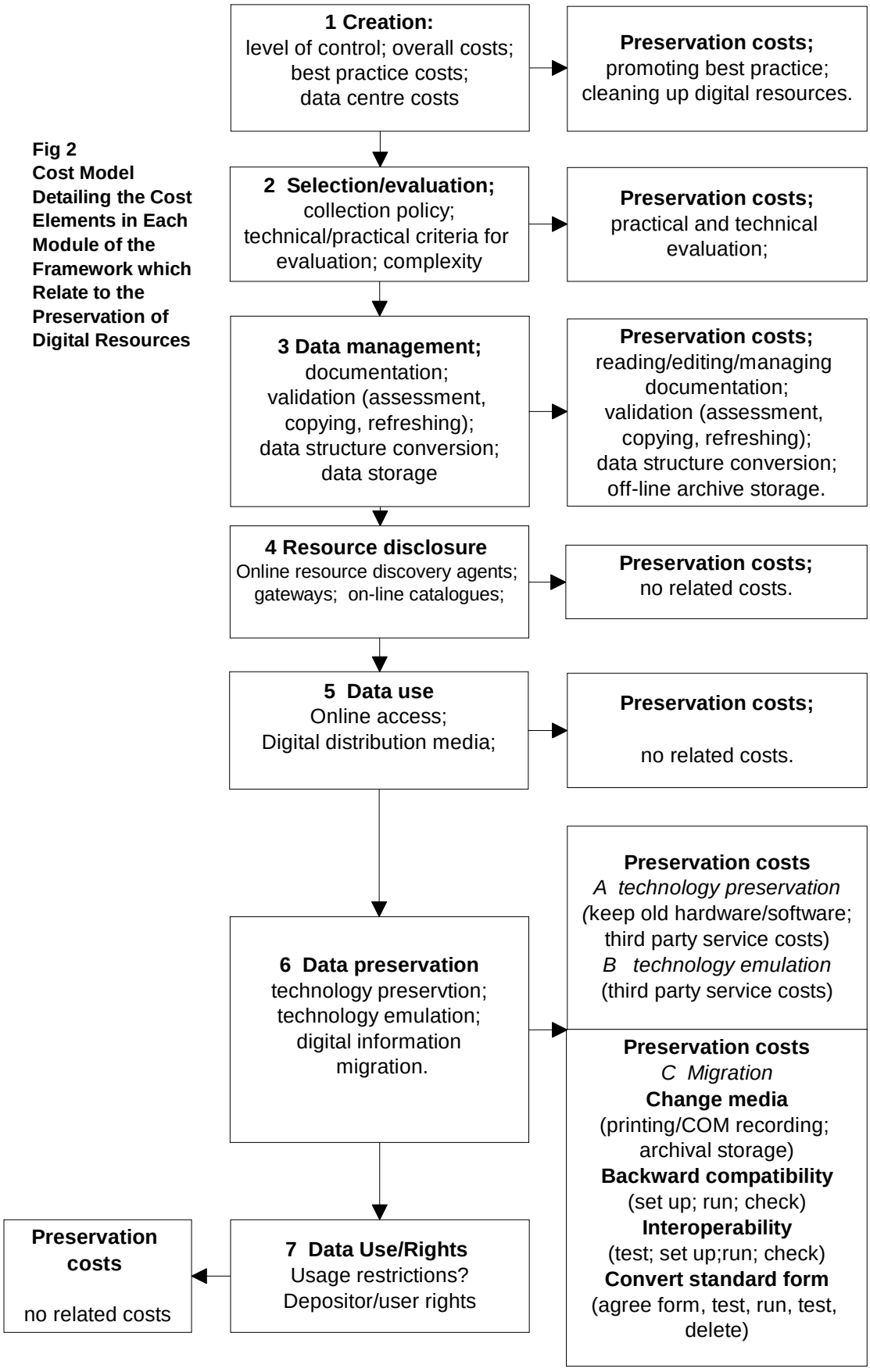
The seven subsections above have defined the tasks involved in implementing each of the seven modules in Greenstein's framework.

The consultancy has identified the key costs, which will be associated with these tasks and has indicated the factors, which will serve to lessen or increase these costs.

Finally the consultancy has indicted which costs are, in its view, directly or indirectly related to preservation and which are not related to preservation.

The resulting cost model is shown in graphic form in Figure 2 below.

Fig 2
Cost Model Detailing the Cost Elements in Each Module of the Framework which Relate to the Preservation of Digital Resources



5.2 Data Sets

The consultancy visited a number of data centres that specialise in managing alphanumeric data sets. These included:

- The Data Archive, Unversity of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ UK. Tel 01206 872 141; 872001(Switchboard); Fax 01206 872003; http://hds.essex.ac.uk/ (Bridget Winstanley, Assistant Director).

- British Oceanographic Data Centre, Proudman Oceanographic Laboratory, Bidston Observatory, Birkenhead, Merseyside L43 7RA UK Tel 0151 653 8633; Fax 0151 652 3950 Email **Error! Reference source not found.** (Dr Meirion T Jones, Director).

5.2.1 Social science and humanities data sets at the Data Archive

The Data Archive is the largest national repository of research data in the social sciences and humanities in the UK. It was set up in 1967 to collect and preserve machine readable data relating to social and economic affairs from academic, commercial and governmental sources and to make these data available for secondary analysis. The work of the Data Archive has been described in several articles and papers (16,17).

5.2.1.1 Creation costs

The Data Archive devote significant resources to promoting good practice to their depositors. They publish a Guide for Depositors on the Web. A significant percentage of the data deposited with them comes from funded research projects and in some of these cases it is a requirement that the data is documented and provided in an acceptable format.

5.2.1.2 Selection & Evaluation costs

The Data Archive has a mandate to collect and preserve certain data. Data is now deposited with them as a matter of course on an ongoing basis i e research data and data from specified Government departments. The Data Archive is also driven by user demand. If their users request certain data they will try and get hold of that data. Storing data sets with all the necessary safeguards and procedures is a labour intensive and hence costly process so they need to be very selective about what they preserve. They envisage becoming more selective in future as there is more data to preserve.

The Data Archive spends a lot of time evaluating digital resources to determine whether they are worth keeping. They check out how well documented each resource is. The cost of documenting data is very high so this has a major impact on their decisions.

The Data Archive takes in approximately 200 new Data Sets per year.

5.2.1.3 Data management costs

The Data Archive check and validate all deposited data. They check that it can be read; it is accurate; it is what they expected it to be and what the Depositor said it was. As part of that they need documentation and as part of that they also check the accuracy and quality of the documentation and note omissions etc.

They carry out a range of consistency checks given the type of statistical data they handle. They also check to see what version the data and documentation is at - have they received an earlier version? Have they received the same version already?

The Data Archive do not document data themselves. If no documentation is sent with the data they will return it and request documentation.

The bulk of the data held at Essex is as follows:

- textual data;
- numeric data;
- alphanumeric data;
- images of documentation.

If they need to deliver data to users online then they also need to deliver the vital accompanying documentation online. This was impossible when all the documentation was held on paper. Hence the Data Archive scanned all the documentation and digitised it. They store the pages as raster images in TIFF file format.

One of the problems they face is that they are dealing with numeric data sets and it is not easy to identify what data you are looking at and where you are in the data set. Hence the documentation is vital and so they justified the conversion exercise.

The Data Archive used to strip out all software dependency from the data they archived when it was received. From the mid 1980s they also began preserving the software that came with the data sets they archived. Now they keep the data sets in the format in which they were supplied to them plus they also process it and save it in a portable (export) format.

They will follow up and test any queries raised by the earlier validation checks and run additional checks if they have any concerns.

All the data sets held by the Data Archive are held on a Hierarchical Storage Management (HSM) system. They use magnetic disk as cache storage for recent and very active data sets. They then have a near-line storage device in the form of a DLT tape library system. This can automatically retrieve any DLT tape and load it into a DLT tape drive, identify the required data set and copy the data across from the DLT drive to the magnetic disk cache ready for usage.

For back-up each data set is shadowed on a separate DLT tape. The Data Archive also employ recordable CDs for back-up and archive and Exabyte tapes. Copies of data sets are written to recordable CD as well as to DLT tapes. In addition, copies of all core data sets are sent to the University of London Computer Centre for archiving (see 5.6 below).

Version control is critical for the Data Archive so they always know where all copies of the same data set are kept and when the data set was last updated etc. They also put a lot of resources into maintaining consistency in processing and storage - to ensure that they treat all their data sets the same way and hold them in the same formats and use the same file labelling conventions etc.

The Data Archive holds approximately 5,000 master Data Sets, which comprise approximately 70,000 Files. The data sets occupy approximately 70 Gbytes of storage. The image files of documentation occupy a further 30 Gbytes.

5.2.1.4 Resource disclosure costs

The Data Archive have built up a comprehensive series of finding aids over the 20 years they have been in existence. They have their own thesaurus. They use AACR top level and their own Authority Lists. Their catalogue is called Biron and is available on the Web. They use 26 broad subject categories so they can pull out all the studies done in any one or more of the 26 subject categories.

5.2.1.5 Data use costs

The Data Archive manages some very large data sets and they have to be able to go in and extract subsets of these large data sets for users. Hence they must be able to extract data and send it to users in the format they want.

The Data Archive issues order forms in addition to their catalogue and indexes. Users can tick which format and media they want the data to be delivered to them in.

The Data Archive will store the data in the most popular formats. If a user requests a more obscure format the Data Archive will convert a copy of the data set into that format on the fly. For the major data sets they will convert them on their server as a batch task. They will retrieve and deliver the relevant documentation with the data sets as well.

5.2.1.6 Data preservation costs

When it comes to preserving data the Data Archive follow the Digital Information Migration strategy. They devote a lot of effort to inspecting and checking and processing the data on receipt to ensure they understand it and can migrate it as required.

There is no hardware dependency on their data. Most of their raw data is held in ASCII form and hence they can process it and migrate it forwards. They process the data to ensure that the data can be processed on more than one piece of software which also gives them some protection.

The Data Archive would not want to rely on just one format. They need a lot of redundancy. They are very much user driven. They are dealing with active data and they have a range of users from novices to very experienced users. Hence one of their prime concerns is to ensure that the data can be used actively and processed using the widest range of software.

They refresh their tape media on a regular basis and they hold 4 sets of each data set. They do not simply rely on backward compatibility of application software. They need all the steps in

the chain so the data is loaded into new versions of the application software and tested.

For the documentation images they have tested the facility to convert from TIFF to Adobe's PDF which gives them a migration path. They need PDF support for Internet delivery.

5.2.1.7 Rights management costs

Contract administration is also a vital area. This involves drawing up a contract with the original data owner and with the users. It defines the access rights and what people can do with the data. It defines what use can be made of the data etc.

All orders for data have to be checked and agreements signed with the users to ensure that they agree to follow the rules governing the usage of the data that were agreed with the owners of that data at deposit time.

5.2.1.8 Overall preservation costs

There are two cost elements. The initial one-off cost of acquisition, cataloguing and processing which is equally shared between dissemination and preservation and therefore has to be halved. The second, recurring cost relates to preserving the dataset over the entire periods for which it is held.

For the Archive, the following formula applies:

Cost element 1

0.5 Acquisition costs+0.5 Cataloguing costs+0.5 Processing costs divided by the number of new datasets acquired in a year.

Cost element 2

Preservation costs divided by the total number of datasets in holdings.

In the case of the Data Archive's current costs for 100 new datasets (approximately the number acquired during an average year) the following would apply:

Cost element 1

50% of the Archive's current annual costs for acquisition, cataloguing and processing is £188,200, therefore:

£188,200/100 = £1,882 per dataset.

Cost element 2

The current annual costs of preservation are £79,200, therefore:

£79,200/500 = £16 per dataset

A dataset held for 10 years would cost:

£1,882 plus £10x£16 = £2,042 over that period.

5.2.2 Oceanographic data sets at the British Oceanographic Data Centre

The British Oceanographic Data Centre (BODC) was set up in 1989. It is the Natural Environment Research Council's designated Data Centre for marine data, tasked with carrying out the following functions:

- Provide data management support for UK marine science;
- Maintain and develop the UK's national oceanographic database;
- Make oceanographic data available to UK academia, industry and government.
- Develop innovative marine data products and digital atlases.
- Collaborate on behalf of the UK in international exchange and management of oceanographic data.

5.2.2.1 Creation costs

BODC started as a generic data centre. It took in data and managed it and made it available. But the problem with this receptive approach was that the quality of the data that was provided to BODC was often not good enough to enable BODC to satisfy the needs of its clients.

Most of the data managed by BODC results from major experiments/ surveys conducted at sea. There are 8 stages involved in Oceanographic data management:

1 Pre cruise planning
2 Pre cruise calibration
3 Data collection at sea + log of activities
4 In situ calibration
5 Data processing, calibration and QC at sea and post cruise
6 Production of publishable data set (fully worked up, Quality Checked labelled and documented)
7 Data as an end product
8 Application independent database (then used to create other databases; secondary usage; products etc)

The problem was that while BODC clearly see the production of good data as an end in itself the scientists involved in the experiments never have the time or resources required to complete steps five and six. As soon as they get data in step 5 they carry out selective processing of parts of the data so they then derive undocumented data and they carry out application processing on subsets of the data of interest to them. The outputs from this activity are the scientific papers.

The result is that often step 6 never happens and the data given to BODC is far from complete or usable by the users of BODC. BODC have to spend a long time processing the data, working it up into something that is usable. If all the steps had been completed in the first place less effort would be needed.

BODC have tried mandating how the data should be processed and documented but scientists are overstretched and do not have the resources to do it. Often the processing is delegated to junior staff and data can be lost.

BODC refer to their old role – that of trying to clean up poor data - as data archaeology. It is frustrating and very expensive work.

Recently they have taken a different approach. They provide a data management service to the cruises where staff from BODC go on the cruise and collect and process and document the data as it is created. This makes it much easier to accept the data and process and package it for distribution and for archiving. They can ensure the integrity of the data.

The cost of the BODC staff is justified when compared to the overall cost of gathering the data, which often equates to £10,000 per day.

One recent big project where they have taken this approach is OMEX 1.

10 man years of BODC staff effort has gone into managing the data at source (while it was being created) and then processing it to produce 2 CD ROMs containing all the data. 2.5 staff spent a total of 4 years each, working on the data for that project.

The overall cost of OMEX was £10 million. The data management cost £300,000 or approx 3% of the total budget. Overall, there are 600 data sets on CD - ROM holding the data gathered on some 47 cruises.

5.2.2.2 Selection & Evaluation costs

BODC products include GEBCO, the general bathymetric chart of the oceans; BOFS, the bio-geo-chemical ocean flux study CD ROM; the North Sea Project CD ROM; UKDMAP, the UK digital marine atlas project; the GLOSS handbook, the GLOSS tide gauge site package and the UK Directory of Marine data sets.

Current BODC projects include OMEX, the ocean margin exchange project; LOIS, the land ocean interaction study; WOCE, the world ocean circulation experiment; EDTEVA, tides editing, visualisation and analysis; NODB, the national oceanographic database and EDMED the european directory of marine environmental data.

As described above, increasingly BODC is involved in the planning stage of new projects and if selected to provide the data management services, BODC staff manage the data gathering and data management processes. Hence the centre is heavily involved in the data creation stage but has less need to carry out selection and evaluation activities as the data has been quality assured and cleaned up before it is deposited.

5.2.2.3 Data management costs

As a general rule, BODC estimate that on a new project some 3 – 5% of the budget goes on data management.

BODC always look to publish the data for a project in an application independent database so it can be widely used and more easily preserved.

For the BOFS data set on CD - ROM they gathered satellite images and RDB tables and produced an accompanying textbook describing the data sets and how to use them. The data in the RDB tables can be loaded into spreadsheets or relational databases and manipulated.

Most of their data is alphanumeric data, which can be loaded into databases. They are now starting to produce electronic documentation using Adobe Acrobat.

Dr Roy Lowry of BODC has published a comprehensive account of the approach taken to data management for the OMEX 1 project (18).

The main data management problem which BODC face is that the data for 50% of the old experiments/cruises has never been worked up so it is archived but it cannot be easily published so it has less value.

It can take BODC five years to work up the data and produce the published data sets for a major project. During that time the data sets are managed online using an Oracle relational database engine and made available via password protection to members of the team.

Oracle tracks the data sets held by BODC and the back up copies etc. The data is all backed up on recordable CDs and DAT tapes.

5.2.2.4 Resource disclosure costs

Data Sets are indexed by geographical coverage. Global; Atlantic; Mediterranean; North Sea; Shelf Seas; Southern Hemisphere; UK Coastal Waters and the Tropics and Gulf of Mexico.

The bulk of the effort goes into structuring the data for each project, organising the data and providing finding aids for the data sets on CD ROM. BODC have moved from being primarily a data archive with some publishing, to being primarily data publishers with a back up archive function.

5.2.2.5 Data use costs

The main method of data distribution is now via the CD ROMs. BODC publish a list of all their data sets on the Web which users can use to order them.

BODC issue order forms in addition to their catalogue and indexes. Users can tick which format and media they want the data to be delivered to them on.

Over the next 5 years they will be producing 6 CD ROMs holding data sets gathered on 100 cruises.

5.2.2.6 Data preservation costs

When it comes to preserving data the British Oceanographic Data Centre follow the Digital Information Migration strategy. They devote a lot of effort to processing and cleaning up and managing the data at the creation stage and on receipt at the centre to ensure they understand it and can migrate it as required.

Part of the processing is designed to make the data application independent. The data sets are loaded into standard relational tables so they can be imported into spreadsheets and a wide range of relational database engines.

When they started they built up 4,000 magnetic tapes holding all their data sets. Now all that data has been transferred to recordable CDs and all new data sets are stored on recordable CDs and on tape.

5.2.2.7 Rights management costs

Contract administration tends to be a more standardised process at BODC. They follow the policies outlined in the NERC Data Policy Handbook (14).

5.2.2.8 Overall preservation costs

In total there are 18 staff at the British Oceanographic Data Centre involved in the activities listed above. Of these the following is an approximate break down of activities derived by the consultancy:

6 staff handle the Promotion Of Best Practice and Support to Cruises (Creation); Selection and Rights Management.

8 staff handle Data management and preservation.

2 staff handle Resource disclosure.

2 staff handle Data use.

The consultancy would estimate that one third of the 18 staff (6) carry out activities, which relate directly to preservation.

5.3 Structured Texts

The consultancy visited one data centre that specialises in managing structured literary texts:

- The Oxford Text Archive (OTA), Oxford University Computing Services, 13 Banbury Road, Oxford, OX2 6NN. Tel 01865 273238 Fax 01865 273275 Email: http://sable.ox.ac.uk/ota/ (Michael Popham, Head of OTA)

5.3.1 Literary Texts at the Oxford Text Archive

The Oxford Text Archive is part of the Humanities Computing Unit (HCU) at Oxford University Computing Services. The HCU comprises the following units:

- Centre for Humanities Computing
- CTI Centre For Textual Studies
- British National Corpus
- Text Encoding Initiative
- Virtual Seminars For The Teaching Of Literature
- Oxford Text Archive

The TEI is a major international research project, the object of which is the production and dissemination of standardised "Guidelines for Electronic Text Encoding and Interchange" (TEI-P3). The European Editor of the Guidelines, Professor Lou Barnard, is also the manager of the HCU and his staff are able to offer expert help on preparing and using TEI-conformant texts.

The Oxford Text Archive (OTA) has been in existence since 1976. Its prime focus initially was to provide an archive where academics creating digital texts could deposit a copy. OTA provided a central repository of digital texts, which other scholars could use.

The coverage was literary works but this had a wide interpretation to include humanities material, philosophical works, legal works etc as well as pure literature. The archive was primarily used by academics until the last 5 years. The texts were deposited in a wide range of languages.

5.3.1.1 Creation costs

OTA devote some resources to promoting good practice to their depositors. OTA issue a Deposit Form, which invites depositors to provide a basic level of documentation for their texts. They have an increasing number of regular depositors who provide texts in the standard formats preferred by OTA.

5.3.1.2 Selection & Evaluation costs

The OTA have produced a Collections Policy (19). Unlike the Data Archive, the OTA is not currently in a position where funding agencies mandate that scholars or researchers deposit digital texts with OTA. In future this may be introduced.

5.3.1.3 Data management costs

Initially texts were accepted in a range of formats and the data was stripped down to ASCII. They would also often keep the original text file – especially if it employed a proprietary encoding scheme. The text data sets were loaded onto magnetic tape and held on a DEC VAX.

Today most electronic texts are deposited via the Web. OTA will look at the text. If it is delivered in a proprietary word processing format they will probably keep it in its original format and also store it as ASCII and will then mark the ASCII version up themselves, as resources allow.

Increasingly texts are being deposited in TEI format. OTA employ the TEI Lite format and will convert tagged text into that format. The Text Encoding Initiative (TEI) is a guideline for using SGML to mark up literary works. There are various versions of TEI for different categories of literary work including plays, poems etc. TEI Lite is a basic standard, which OTA follow. OTA now have a significant number of regular depositors who observe the preferred standard.

OTA do not receive much material in HTML format and they do not employ HTML as an archive format. This is because it only supports a small number of tags and hence is not as flexible as TEI Lite. Also, HTML does not provide a standard header.

Increasingly OTA are looking at the new XML standard. In order to support XML fully then the literary world will have to rewrite TEI to support both SGML and XML. It is much easier to write software for XML than for SGML. Compared to HTML then XML will be much more flexible. HTML is tied to a limited number of tags. XML is extensible.

OTA will look at the text and if it has not been tagged they will produce a basic Header so they can create an index record for the text and provide a basic catalogue record for the text.

The next step would be to load the text files. OTA now make use of the University's HFS storage system. HFS comprises two IBM RS6000/R40 processors with a total of 1024 MB RAM and 300 GB hard disk storge. They have a 3494 automated tape library with a storage capacity of 1200 Megastar 3590 tapes i e 12 TeraBytes or up to 24 TBytes if data compression is employed.

All the data is backed up on the system. For contingency OTA will hold the TEI Lite version of the text plus either the original text or an ASCII copy. Both SGML and XML rely solely on ASCII character text so there is no proprietary data held.

OTA currently hold approximately 3,000 data sets – some of which comprise large text collections themselves. They estimate that this comprises a total of approximately 5 Gbytes of data.

In resource terms they estimate that it takes them 10 minutes to load each text file and give it a basic header; up to 2 hours to process the data initially and catalogue it etc. It can then take from 1 day to one week to mark up the data, depending on the size and complexity of the text.

5.3.1.4 Resource disclosure costs

Each text loaded into the archive must have a Title and then they will enter additional data into a range of fields including a Source Description; Keywords; Text Classes; Language Codes etc.

They are moving towards the production of MARC records for all electronic texts so their records can be held on the University Library's on-line catalogue.

They will follow the lead set by AHDS who are developing standards for a centralised catalogue. The requirements are reviewed in a comprehensive report edited by Paul Miller and Daniel Greenstein (20).

5.3.1.5 Data use costs

Initially, most requests received were from other academic institutions. The OTA satisfied these requests by copying a block of text files to tape and shipping the tapes out.

The widespread introduction of PCs and the Web has changed the way in which scholars request texts and hence the way in which the OTA provides texts.

At present OTA hold the texts in an archive format. If researchers wish to view the texts then they have to use application software to process the data and render it. OTA list recommended software including SGML browsers, parsers and editors.

OTA are developing a replacement Web site and an online catalogue. This will allow users to search for all the texts held by OTA and access the header data. The texts will be converted automatically via converters from TEI Lite to HTML or ASCII or RTF or users will be able to access the SGML text via style sheets.

At present there is no direct route from SGML to Adobe's Acrobat PDF – OTA recommend going from SGML to RTF and then from RTF to PDF.

When XML becomes widely supported then OTA could provide a simple style sheet from XML and use that to deliver texts.

Currently they only have a small subset of their texts available on the Web but they still receive approximately 5000 hits per month. They receive approximately 1,000 requests per year to access texts offline.

OTA are also looking at the use of DSSSL for publishing documents in TEI Lite.

5.3.1.6 Data preservation costs

When it comes to preserving data the Oxford Text Archive follow the Digital Information Migration strategy. They devote a lot of effort to processing and marking up the texts on receipt at the centre to ensure they understand them and can migrate them as required.

Part of the processing is designed to make the text data application independent. If it is delivered in a proprietary word processing format they will also convert it to ASCII and will then mark it up themselves, as resources allow.

5.3.1.7 Rights management costs

Contract administration has tended to be a more standardised process at OTA. They anticipate devoting more resources to this area in future.

5.3.1.8 Overall preservation costs

In total there are 3 staff at the Oxford Text Archive involved in the activities listed above. In addition they make use of the University's computing centre for data storage and management services.

The following is an approximate break down of activities derived by the consultancy:

0.5 staff handle the Promotion Of Best Practice (Creation); Selection and Rights Management.

1 staff member handles Data management and preservation.

1 staff member handles Resource disclosure.

0.5 staff handle Data use.

The consultancy would estimate that half of the 3 staff (1.5) carry out activities, which relate directly or indirectly to preservation.

5.4 Office Documents

Most of the data centres visited by the consultancy manage and preserve some "office documents" as part of their collections. However, in most cases "office documents" are not their prime focus.

The long - term management of digital "office documents" as electronic records is the prime concern of the Public Record Office's EROS Programme. The long term management of digital office documents is also the concern of many corporate archives and electronic record centres. Increasingly all public bodies will need to manage digital "office documents" as electronic records.

The consultancy has assisted many public and private sector bodies in implementing electronic document management systems to manage their active and inactive digital office documents. Increasingly, these organisations want to manage all their active documents in an electronic document management system and then to carry out standard records management reviews and preserve their core digital office documents as electronic records.

Most organisations are still in the process of defining their electronic records management requirements and procedures. In this section, the consultancy has opted not to report on actual practices and costs in one or two centres. Rather, the consultancy has undertaken a broader survey of current and planned practices. Based on that broader survey, the consultancy has taken the seven cost categories used in the cost model and reviewed all the main costs involved in preserving the full range of digital office documents as electronic records.

5.4.1 Creation costs

The Public Record Office (PRO) has statutory duties to provide guidance and supervision on the management and selection of electronic records used throughout government as well as the more traditional paper records.

The PRO EROS (Electronic Records in Office Systems) Programme issued a "Checklist of Good Practice for Electronic Records Management". They have also now issued a draft set of "Guidelines on the Management of Electronic Records" (6) for comment.

The guidelines are designed to enable government departments to achieve the good practice defined in the checklist. The guidelines aim to describe what government organisations need to do to achieve records management and archiving facilities in the systems which are used to create, access and preserve their electronic documents. The guidelines will do this by identifying the requirements and demonstrating alternative strategies for achieving suitable facilities. The guidelines also provide advice, guidance and technical instructions to departments to enable them to transfer records to the PRO in electronic form.

The guidelines were aimed at departmental business managers, departmental records officers (DROs), heads of IT, strategy and planning managers and PRO Inspection and Documentation Officers (IDOs). The IDOs are the PRO staff who work with government departments to select a subset of the records, which have passed the second review process by the Department, for permanent deposit with PRO. Increasingly these will be digital office documents.

Clearly the production of the guidelines and the liaison required with all the government departments is a very significant undertaking and the PRO have a subset of their Government Services Division dedicated to this task. They are publishing case studies of departments who follow good practice in this area and they run courses and training sessions for government department as well.

The same proportionate amount of effort is required within a single government department or a large corporation seeking to introduce a corporate – wide electronic document management and electronic records management system. Current paper based records management procedures have to be reviewed and modified to cater for the requirements of electronic records management.

The PRO guidelines cover the following areas:

- Creation and use of electronic records;
- Disposal policy;
- Appraisal of electronic records;
- Preservation of electronic records;
- Transfer of electronic records to PRO

The PRO guidelines define Electronic Document Management as follows:

> "Electronic Document Management uses applications to manage the storage, versioning, searching, retrieval and contextual information of electronic documents, either images or electronically authored documents. Unless the EDM application has specific, deliberate ERM functionality added, it is unlikely to deliver records management functionality".

The PRO guidelines define Electronic Records Management as follows:

> "Electronic Records Management provides an electronic framework for capturing electronic documents and applying standard records management practices. This functionality incorporates a corporate filing structure, document classification within the filing structure and formal retention and disposition scheduling based on an approved disposition and review schedule."

The PRO are devoting considerable resources into the promotion of best practice in government departments. All enlightened organisations are devoting similar resources to the promotion of best practice in electronic records management. They all recognise, like BODC and OTA and the Data Archive in their fields, that good practice at the creation stage can avoid huge costs at the data management and preservation stages.

5.4.2 Selection and evaluation costs

The PRO shares the costs of selection with Government Departments. Currently departments carry out a first review of their records and destroy ephemeral records etc. They then hold them for a further 15 years and carry out a second review to weed out a large proportion of the records. The remaining core records are held until they are due for deposit with the PRO (25 – 30 years). The PRO Inspection and Documentation Officers (IDOs) then inspect those records and select a small percentage of them for permanent storage at the PRO.

The figure stored at PRO represents between 1 and 3% of the total records of each

department, on average. The bulk of the selected records are made available to the public at the PRO after 30 years. That figure may reduce in future with the moves to promote the freedom of information.

In future, when the majority of these records are electronic and digital storage costs are insignificant it could be argued that the amount of effort devoted to selection could be reduced. However, it is expected that some level of review and selection procedures will always be needed simply because it is not good practice to keep records longer than needed.

The evaluation process used to involve inspecting paper records to see whether they were legible, whether they had been filed correctly and whether their format required the use of special storage equipment (large format records etc).

With electronic records, in future, the PRO will be checking records against technical criteria. The PRO guidelines lay down a preferred list of transfer formats and the PRO will have to check all deposited electronic records to ensure they have been transferred in one or other of these formats. The draft guidelines define the requirements for a transfer format:

> "The transfer formats need to be comprehensive in that not only document content but structure and context must be part of the archived document. At the same time transfers need to be made with minimal interference to departmental work loads and little, if any, requirement for extra software to be installed on the system supporting the assembly to be archived."

The standards, which the PRO recommend for transfer in the draft guidelines are – PostScript; TIFF, SGML, Adobe's PDF and delimited file format (comma separated variable).

5.4.3 Data management costs

The PRO, or any large corporate records centre, must validate the electronic records (digital office documents) transferred to them. They will require the creation of finding aids and record metadata for each class of digital office documents transferred to them. The PRO Guidelines contain details of a draft Document Profile.

It will be the responsibility of the Department to prepare appropriate finding aids. The PRO will archive the finding aid with the electronic record. The PRO Inspection and Documentation Officers would advise departments on the most appropriate structure and layout for the finding aids.

Corporate electronic records centres will also have to draw up standards for metadata and indicate which fields must be completed and which are optional.

The next step, after the records have been accessioned and validated, will be to store them. The PRO defined the primary objectives of the storage formats as being:

> "the preservation of record assembly integrity and, as far as is possible, stability. There should not be too many formats and they should be managed in such a way as to inspire confidence in their longevity. One important aspect of the preservation formats is that they should be capable of easy conversion into a presentation format."

At present the PRO have made the pragmatic decision that the preferred standard formats identified for transfer are also suitable for preservation within the PRO.
The PRO have reviewed a range of storage media and they currently prefer the use of write

once recordable CDs or CD ROMs.

Many corporate electronic record centres will want to manage electronic records online for ease of access and off line for back-up and archival purposes. They will be in exactly the same position as the data centres described above. Good practice will dictate that copies of all the electronic records should be made and the archive media should be stored off - site away from the active storage systems. They may opt to use third parties for the archive storage function or even for the active on-line storage (see 5.6 below).

5.4.4 Resource disclosure costs

As indicated above, the PRO will require Departments to provide suitable finding aids. The PRO are in the process of implementing an online catalogue that will be made available via the Web.

Corporate electronic record centres would define standards for metadata and indicate which fields are mandatory and which are not. The index data for all electronic records would be held online.

5.4.5 Data use costs

Increasingly, core digital office documents will be converted into standard formats, such as those recommended by the PRO, for long term management as electronic records. Where the level of usage justifies it they will be held online and made available via gateways and Internet browsers. For preservation purposes they will be archived onto multiple copies of the preferred media (recordable CDs; DLT tape; DAT tape etc) and held off site, possibly by third party service providers (see section 5.6 below).

If the depositors can be persuaded to provide the documents in the preferred formats then the costs of holding them in the preferred formats and rendering them into the required presentation formats should not be excessive.

5.4.6 Data preservation costs

The PRO, in their guidelines, and all corporate electronic records management centres the consultancy has visited, have adopted the “Digital Information Migration” strategy. The preferred subset, as indicated above is “Convert to standard formats” with “Change media” as a backup strategy for the short to medium term. The back up strategy for record centres would be to record the document page images onto microfilm.

The PRO guidelines make the following points:

> “the rate of change associated with the development of applications and hardware platforms is likely to shape the timetable for migrating electronic records. Should that not be the case these records should be copied on to new media at intervals that meet the manufacturers recommendations for the medium to prevent the physical loss of data or technological obsolescence. There is no fixed frequency or interval for when this need may arise. The DRO will need to liase with the IT department to identify when this is likely to occur and to ensure that appropriate procedures are in place to safeguard both the records and the associated metadata. It would be unusual if migration occurred more frequently than three years.”

5.4.7 Rights management costs

The main “rights” issues for the PRO would relate to the confidentiality of electronic records deposited with them prior to them being released to the public after 30 years. They would have to ensure that such records could not be viewed by the public prior to them being released.

Secondly, the PRO would have to consider whether they needed to impose restrictions on the volume of data which members of the public could download from their system and hold locally.

For corporate electronic document and records management centres to function efficiently then certain policies and principles would need to be agreed. The first one would be that all information/records created by staff while working for the organisation would be the property of the organisation. Hence they should be made available to all other staff on a “need to know” basis and individuals could not claim copyright.

The electronic records centre manager would have to ensure that no third party material was held on the system without the express permission of the owner of the copyright or without the purchase of the required license or payment of the required royalty charges.

Reaching consensus on these points and checking that they are being complied with will represent a significant cost for every large organisation.

5.4.8 Overall preservation costs

For the costing exercise, the consultancy shall take, as an example, a generic head office of an organisation with 500 staff and 200,000 digital office documents stored in the electronic records centre as core long term electronic records.

The main costs which that organisation would have incurred, initially, would have been in upgrading their IT infrastructure to support electronic document and records management and in scanning and digitising all their new incoming paper documents and their active existing paper documents.

The IT infrastructure costs would include upgrading the network to provide at least 10 mbs and 100mbs on backbones. It would include providing all staff with high specification PCs with sufficient memory and high resolution displays to view document images. It would include providing sufficient mass storage on the network servers to manage all the active digital documents online on magnetic storage and providing near-line storage for holding inactive digital documents and back-up copies of all digital documents. It would also include the cost of the licenses needed for the database engines and the required electronic document management and electronic records management software.

It would also include the costs involved in defining the requirements for electronic document and records management and in developing the required system and in training all the staff to make full use of the facilities.

Depending on the amount of infrastructure upgrades required the total costs of implementing such a system would range from £2,000 to £5,000 per user.

The yearly running costs for the system once implemented would average £1,000 - £1500 per

user. The costs attributable to preservation would usually represent approximately 35 - 40% of the running costs. They would cover the staff resources required to promote good practice amongst all staff and to train staff in following the required procedures. They would cover the staff resources involved in accessioning records into the long - term electronic records storage system. They would include staff costs involved in checking and correcting index data and validating records. They would include staff costs involved in copying and refreshing data and in migrating data into new preferred formats. They would involve the cost of archiving data off-site with a third - party commercial storage company. Examples of the costs of third party data archive services are provided in section (5.6) below.

Approximately 5 dedicated staff would, typically, be needed to preserve and manage the long - term electronic records in the above example.

5.5 Visual Images

In this section the consultancy has undertaken a broad survey of current and planned practices. Based on that broad survey, the consultancy has taken the seven cost categories used in the cost model and reviewed all the main costs involved in preserving the full range of digital visual images including scanned images of manuscripts and bound volumes.

5.5.1 Creation costs

As described in section (4.8) above, at the creation stage there are two main issues, which have to be considered carefully by data centre managers when preserving visual images.

The first issue relates to the resolution at which the images should be scanned and whether they should be held as black and white, greyscale or colour images. The options are listed in section (3.4.1).

The second issue then relates to what file formats the image data should be stored and interchanged in – whether or not the image data should be compressed and, if so, what compression algorithm should be used. Again, the main options are listed in section (3.4.1).

5.5.1.1 Capture quality

All data centres managing and preserving digital images should devote some resource to the promotion of good practice. They should visit centres that have captured a range of digital images and gain from their experience. They should study the literature and, most importantly, they should study the type of images that their users are depositing with them. They should then issue guidelines to depositors and those guidelines should indicate the preferred range of resolutions and coding schemes for capturing specific categories of images.

A considerable volume of literature has been produced on this subject (8 – 12). It represents the results of millions of dollars of research. It also serves to indicate that this is not a simple issue. The consultancy has produced a select bibliography in the appendix to supplement the above references and recommends that readers interested in the many issues relating to the digitisation of a wide range of image material, should follow up some of the references.

The choice of resolution and coding schemes has a very significant impact on the cost of the capture operation and on the cost of subsequently managing and preserving the digital images.

A significant proportion of data centres and digital libraries will also face the costs of actually scanning and digitising images if the images are not held in digital form at the time of deposit and the policy of the centre/library is to manage their images in digital format.

The big policy issue, which the centre has to decide on in this situation, is why are they digitising the images? There are usually three answers to this question.

- to improve access to them;
- to preserve them and avoid having to handle the originals again;
- to both improve access to them and preserve them.

The answer to this question will determine whether or not they regard scanning and digitisation as a one - time exercise – this is the only opportunity they have to capture the image – or whether they are happy to contemplate repeating the exercise in 5 or 10 years time.

The articles in the bibliography explore this question in far more detail than the consultancy can in this study. However, a few key points can be made to see why the answer given to this question can have such an impact on the costs.

A Improve access only

If the centre is only digitising the images to improve access to them then this implies that they will be running a hybrid system. The original analogue images will still be available for users who request the originals. The purpose of digitisation will be to provide users with fast access to surrogate images on their computer screens to aid them in deciding which images they are interested in and to meet a certain level of reference requests. In such cases, a compromise could be made on resolution and the levels of coding needed. There may be little point in capturing the images at a resolution higher than that supported on most computer displays or printers.

B Preserve images only

At the other extreme, the centre may be digitising the images purely for preservation purposes. They may have a one - time opportunity to capture the images. This may be because the images are deteriorating and cannot be handled again or because they belong to a private collection and are only being made available for scanning once or because, pragmatically, the centre has a budget to do the digitisation and cannot guarantee that they will ever get the budget again. In this case, the purpose of digitisation will be to capture all the data held on the original image at the highest quality affordable so that, for all foreseeable purposes in the future, the digital image can be used as the master from which distribution copies etc will be derived. In this case, the only reason for compromising on quality would be budgetary reasons.

C Improve access and preserve images

The compromise position is that the centre is digitising the images to preserve them and improve access to them. This is by far the most common situation. The ideal would be to scan the original images at a sufficient quality so that there is no need to go back and rescan them again. The resulting high - resolution images would be held on the archive storage media and refreshed and migrated at regular intervals for preservation. For access then, for as long as there are network bandwidth, screen resolution and printer resolution limitations, then pragmatic decisions will be made to also hold the images at reduced resolution and size on the active management system or to convert the master images to lower resolution images in real time prior to sending them across the network.

If the available images are held in certain formats then there can be limits imposed on the resolution they can be scanned at using standard equipment. At the time of researching this report the consultancy could not identify any production microfilm scanner that could scan at a resolution above 400dpi (dots per inch) without custom modifications. Similarly, there were no book scanners in production that could scan above 400dpi although a 600dpi book scanner is promised.

The US National Archives and Records Administration (NARA) have set up an "Electronic Access Project" (EAP) which is designed to produce an online catalogue that will provide information about NARA holdings and a core collection of digital copies of selected high interest documents (21). EAP will provide online access but according to NARA "will not address preservation issues using digital technology at this time". This puts EAP into the category (A) above. NARA regards this project as a pilot and they have gone for a high scanning specification. All scanning will be done at either 8 bit greyscale or 24 bit colour.

For textual documents including the originals, photocopies and microfilm they are scanning at two resolutions. 300 dpi will be used for documents up to 11 x 17 inches where they are going to apply recognition technology to the resulting image to recognise and code the text. 200dpi will be used for larger documents to be of reproduction quality and to save file storage space.

The photographic material includes black and white and colour photographic prints, negatives and transparencies. NARA will use a digital image size for the master files of 3,000 pixels across the long dimension by the proportional number of pixels for the specific photo format (3000 x 2400 pixels for 8 x 10 inch prints; 3000 x 2000 pixel for 35mm slides). NARA claim that this provides reproduction quality as a magazine quality halftone reproduction of 11 x 14 inches in size and 133 lpi is achievable from this file size. Final image size will be set to a standard 10 inches across the long dimension at 300 dpi.

The Electronic Text Centre at the University of Virginia has produced a helpful guide to image scanning (8). For archival imaging (B) they recommend a default scanning resolution of 400 dpi. They also recommend that centres scan at 24 bit colour by default. They recommend the use of the JPEG compression algorithm because the resulting compressed JPEG file from a 24 bit original will be smaller than that made from an 8 bit original.

The Library of Congress have published a number of documents designed to present a snapshot of their digital conversion activity in the "American Memory" pilot programme and the operational "National Digital Library" programme. One interesting article, by Carl Fleischhauer (22) covers "Digital formats for content reproductions".

Image Type	Tonal Depth	Format	Compression	Spatial Resolution
Thumbnail	8 bits per pixel	GIF	Native to GIF	Circa 200 x 200 pixels
Reference	Greyscale : 8 bits per pixel; colour; 24 bits per pixel	JFIF (JPEG File Interchange Format)	JPEG (generally about 10:1 compression)	Moderate class ranges from about 500 x 400 to 1000x 700 pixels; Higher resolution class (future) will range from 2000 x 1400 to 4000 x 3000; both moderate and higher resolution will be offered to users
Archive	Greyscale : 8 bits per pixel; colour; 24 bits per pixel	TIFF (Tagged Image File Format)	Uncompressed	Moderate class ranges from 500 x 400 to 1000 x 700 pixels; higher resolution class will range from 2000 x 1400 to 4000 x 3000; only the highest resolution will be archived

For pictorial collections the Library produce three image types. The table above contains the specifications for each type. Hence the Library of Congress are covering all three options (A – C) in their project.

Yale University Library, together with Xerox Corporation have been working on "Project Open Book" to scan and digitise the content of volumes which had been microfilmed in the past. The decision was taken to scan the microfilm images rather than the original books. In his report on the set up stage of the project (23) Paul Conway, head of the preservation department at Yale University Library describes the results of their investigation into image quality issues. The project was broken down into three stages. The first stage covered the set-up and the digitisation of 100 volumes. The second phase was to cover the production scanning of the next 3,000 volumes and the final phase would cover the production scanning of a further 6,900 volumes to make a total of 10,000 volumes.

A Mekel M400 microfilm scanner was used in the set up phase for 35mm film scanning and 100 volumes were scanned. The volumes were stored as black and white images. Yale were able to obtain a scanning resolution equivalent to 600dpi from the Mekel scanner after some customisation by a supplier called Amitech. The maximum capacity of the CCD array of the Mekel scanner was 3694 pixels per inch at the full 35mm width. The conversion of microfilm at 600 dpi is a process of adjusting the capacity of the CCD array in relation to the reduction ratio of the material preserved on film. 600 dpi is therefore a software-controlled mathematical artefact of the scanning process.

In his report on the production phase of the project (24) Paul Conway reported that 2,000 volumes were actually scanned in a 12 month period. He concluded that high resolution scanning of at least 400 dpi is essential for preservation quality digital conversion of text. He also concluded that high quality binary scanning produces high quality digital images of text. In other words he does not agree with some of the other projects above that greyscale scanning is essential even for text documents. A third firm conclusion was that high quality digital images can be generated from film created within the standards and guidelines for archival quality preservation microfilm. Conway also confirmed that in order to obtain 600 dpi resolution Amitech Corporation redesigned the Mekel 400XL and added special software.

Conway defined the capture cost as totalling $55.03 per book or 25.4 cents per image. This assumes an average of 217 images per book. Given that this involved scanning existing microfilm and only involved capturing black and white rather than greyscale or colour images, this gives an indication of the high costs involved in capturing high volumes of high quality digital images.

The consultancy conducts an annual survey of scanning bureaux to obtain average service costs for the scanning of office documents (5). The following figures were obtained for 1998.

Size	Medium/Format	Resolution	Price	Volume
A4	Single page side	200dpi	6.12p	Per page
A3	Single page side	200dpi	8.38p	Per page
A2	Drawing sheet	200dpi	8.38p	Per page
A1	Drawing sheet	200dpi	10p	Per page
A0	Drawing sheet	200dpi	13.4p	Per page
16mm	Roll microfilm	200dpi	8.9p	Per frame
35mm	Roll microfilm	200dpi	20.5p	Per frame
	Microfiche	200dpi	9p	Per frame
35mm	Aperture card	200dpi	61p	Per frame/card

These figures cover the scanning of office documents at a low resolution and bitonal capture. Once the requirement moves beyond that to scanning a range of material at resolutions up to 600dpi and capturing 8 bit greyscale or 24 bit colour then the costs rise very significantly.

Another factor, which affects the costs significantly, is the quality of the original and whether they are bound (books) or mounted (slides, transparencies etc).

If the requirement is for 400dpi resolution, greyscale images of the pages of a bound volume then this implies the use of a book scanner with greyscale capabilities. The per image costs here will range from £1 - £5 depending on the quality of the original and the amount of handling and image processing required.

If the requirement is for 600dpi resolution, 24 bit colour images then this implies the use of a flat bed scanner. The per image costs here could range from £3 to £10 depending on the quality of the original and the amount of handling and image processing required.

5.5.2 Selection and evaluation costs

Given the high costs involved in capturing and processing digital images, considerable care needs to be given to the selection process. This, in turn, means that the costs associated with selection are also very high. The National Archives and the Library of Congress projects described above would have involved considerable staff resources at the selection stage. The Project Open Book also involved considerable work at the selection stage, which would represent an additional cost over the capture costs quoted above. A centre that just wants to scan all the images in one collection, will still have to carry out all the following tasks:

- review the entire collection;
- check and list the contents;
- describe the range of material contained in the collection;
- plan the sequence in which the material should be scanned;
- handle all the material.

If the centre has carried out the scanning and digitisation exercise then they would not incur any additional significant costs evaluating the material prior to loading it into the collection.

However, if a large collection of digital images is deposited at a centre then a considerable amount of work will be involved in evaluating the resource. The Visual Arts Data Service, part of the Arts & Humanities Data Service, have produced draft Guidelines for Depositors (25). This includes a series of criteria that they will use when evaluating data sets.

5.5.3 Data management costs

Jennifer Trant provided an excellent account of good practice for documenting digital images in her paper, presented at the Museum Computer Network Joint Conference (11).

She makes the point that ideally the depositor should record information both about the original and about the digital representation of the original because the characteristics of the original will often have an impact on the resulting quality of the digital image. The same is true if the image was derived from or subsampled from another digital image file.

She covers the documentation required to describe the content, the surrogate images and the digital image files. She also covers standard means of representing the various types of image descriptions.

If a data centre receives the bulk of their digital images from a few major depositors then they should be able to agree standards for documentation and hence the costs to the centre will be small. Where the centre has to check and edit documentation prior to deposit then they would incur very significant costs.

Validation of digital images would occupy considerable resources if centres opted to check every image – particularly if the centre was storing high - resolution colour images of large originals. Clearly if the images all belonged to one book, as in the Yale project, then the overhead involved in validating them would be less than if they were all unique images. With the Yale project there would be scope for selective checking and consistency checking.

If the centre does the scanning and digitising then they would save the images in the required file formats and using the required compression algorithms. Hence there should be no need to alter the structure of the digital images prior to storage apart from deriving thumbnail or reference images from the archive image.

If the centre receives the digital images as a deposit then they my not be in the preferred formats in which case the centre would need to import them into a preferred image processing application and then convert them into the required format and save them in that format. Again this could involve significant resources when dealing with high volumes of large image files.

Finally, the centre will have the cost of storing the digital images in one or two or three different formats. The costs can be divided into active online storage costs and off line back up and archive storage costs.

The online storage costs will cover the cost of storing the thumbnail and reference images – using the Library of Congress model. Assuming 24 bit colour images and JPEG compression averaging 10:1 the file sizes required for the spatial resolutions described by the Library range from 60 Kbytes to 210 Kbytes for moderate resolution to 840 Kbytes to 3.6 Mbytes for higher resolution.

If we assume equal quantities of each image in a large image database then we reach an average file size of 1.18 Mbytes. Hence an online database of 100,000 images would require approximately 118 Gbytes of online storage. An online database of 1 million images would require approximately 1.18 Tbytes of online storage.

The off-line costs for archival storage of large high - resolution image databases can also begin to get significant. The highest resolution archive images, stored uncompressed, will total 36 Mbytes per image so an archive containing 100,000 images would require 3.6 Tbytes of off-line storage media. An archive containing 1 million images would require 36 Tbytes of offline storage media. Three copies would need to be kept so the total media requirements would exceed 100 Tbytes.

The copying and refreshing costs associated with that volume of data would be very significant.

5.5.4 Resource disclosure costs

The Library of Congress have published a number of documents designed to present a snapshot of their digital conversion activity in the "American Memory" pilot programme and the operational "National Digital Library" programme. One interesting article, by Caroline R Arms (26) covers "Access aids and interoperability". It looks at the identifiers required for digital reproductions and suggests a number of formats for access aids including bibliographic records that adhere to the MARC format and simple bibliographic records following the emergent Dublin Core approach.

In the AHDS report (20), Catherine Grout of the Visual Arts Data Service contributed a section summarising the findings of a workshop examining the descriptive information needed to enable the discovery of visual arts, museums and cultural heritage resources on the Internet, particularly in the form of digital images. A full account of the result of the workshop authored by Catherine Grout and Tony Gill is available on the Visual Arts Data Service site on the Web (25).

Once standards have been agreed then if regular depositors provide the required bibliographic records the main costs incurred by the centres will relate to checking the data and making it available online. The costs of creating the required resource discovery data at the centres would be extremely high if they were receiving high volumes of digital images.

5.5.5 Data use costs

The main costs associated with providing users with online access to digital images in thumbnail and reference format, will be the costs of running the Web server/s and the cost of providing dedicated workstations and printers for users to use at the centre - if that service is provided. Because the level of access is hard to predict, many centres will contract the service out to a third party to manage.

5.5.6 Data preservation costs

All digital libraries and archives managing digital images have either adopted the "Digital Information Migration" strategy or else they are still preserving the original images and simply regard the digital images as surrogates for online access.

The preferred subsets of the "Digital information migration" strategy are "Convert to standard formats" with "change media" as a back up strategy.

The main costs, in addition to the off line storage and refreshing costs referred to in (5.5.3) above will relate to the costs of migrating digital images from one non preferred format to a new preferred format as required. As more experience is gained in this area so more tools and macros could be used to automate part of this process. Selective testing and checking will still be required and will be labour intensive in major digital image archives.

5.5.7 Rights management costs

Rights management would be a major issue for collection managers looking after large digital image collections. Hence the staff resources required in this area would be significantly higher than those required by centres looking after data sets or office documents or even structured texts.

5.5.8 Overall preservation costs

The first point to make here is that it would seem unlikely that a centre would undertake the large-scale scanning and digitisation of visual images purely for preservation purposes. The majority of the projects that are underway today, some of which are described above, have been driven by the desire to improve access to a selected set of images. Hence most centres in this area will be looking to depositors to provide them with digital images to preserve.

There will be significant staff costs involved in promoting good practice for depositors, selecting digital images for the collection and evaluating those offered for deposit.

If the centre is to provide online access to reference copies of the digital images then, if the collection is likely to grow to 100,000 images or more there will be a significant cost involved in managing that volume of data and all the associated attribute data online. The centre should consider the option of a third party service provider carefully. These costs do not strictly constitute preservation costs.

Unless standards for index records or metadata can be agreed between the centre and the depositor the centre face very significant staff costs preparing such data for high volumes of images.

Provided the centre can afford to mount all their reference and thumbnail images online then the cost of making the data available to users should not be excessive. If the centre opts to provide a significant number of user workstations and high quality colour printers at their centre for users then the costs would increase significantly and could average approximately £5,000 per user.

The main preservation costs will be the cost of storing the high - resolution archive images uncompressed off-line and refreshing and migrating the data as required. Again- such a service can be outsourced and some of the options and costs relating to third party archiving are looked at in the next section of the report (see 5.6 below).

There are really too many variables to arrive at any overall generic cost figures for the preservation of digital images. How many images per year will come in; how much documentation and cataloguing will be required; what will the level of accesses be and how often will the digital images need to be migrated. Will the archive data be managed by a third party?

Based on current experiences and a knowledge of similar environments, the consultancy would estimate that for a digital image archive holding 100,000 digital images and receiving up to 10,000 new digital images per year, the preservation tasks would require two to four dedicated staff posts.

5.6 Commercial data storage costs

All the data centres/archives/libraries, irrespective of the type of digital resources they manage, share a common need to create archive copies of the data and hold them offsite in controlled storage conditions. Increasingly, they also share a requirement to make the digital resources available online.

5.6.1 Off-site archiving with commercial data archive companies

For off-site archiving, commercial companies such as Hays Information Management and Britannia Data Management will hold archive copies of the data for a client at their premises.

They will charge a fee for cataloguing each tape or CD etc onto their records management system and producing a bar code label etc.

They will charge a monthly or yearly fee for holding each tape or CD etc.

They will charge a fee for any media handling requirement (retrieval; copying; refresh etc)

If off – line access to the media is required, they will charge a fee for transporting each tape or CD back to the customer.

For preservation purposes, if a data centre wanted to send 2 copies of each tape/CD offsite then they would pay for initial handling, ongoing storage and ongoing handling. If the centre had a disaster on their active server they would pay for the relevant tapes/ CDs to be transported back to them and the data reloaded.

A data centre managing a total of 200 Gbytes of data sets would send 2 copies of all media to the data storage company.

If they used a high density storage medium such as DLT tape they could hold 40 Gbytes or more on one tape. However, given this would mean storing thousands of data sets on one medium they would be unlikely to hold more than 10 Gbytes on each cartridge.

Hence they would send approximately 20 x 2 = 40 cartridges to the storage company.

The offline storage cost plus the initial handling cost per cartridge would be £6 – 7 per year so the total yearly storage cost would be just £240 per year. There may be a minimum administration cost to add to this.

A centre with a large collection of digital images comprising 36 Tbytes (see 5.5 above) would be storing 72 Tbytes on a maximum total of 7,200 DLT tapes.

The offline storage cost plus the initial handling cost per cartridge would be £6 – 7 per year so the total yearly storage cost would be £50,400 per year.

Beyond that then, for an additional range of costs, the commercial storage companies could provide a proactive refresh programme and could also provide similar on-line and near-line storage services to those described below from ULCC.

5.6.2 Near-line Storage from ULCC.

A number of University computer centres such as the University of London Computer Centre (ULCC) offer data archiving and near-line data storage services. ULCC has established the "National Data Repository Service" which allows users to access and search enormous volumes of data as if it were on disk, while maintaining the cost advantages of tape storage.

The total capacity is potentially up to 300 Terabytes. Once data is within the system, ULCC takes responsibility for ensuring that it is managed safely. They hold two copies of all files and tapes are "sniffed regularly" to guard against media degradation. As tapes age, files are moved automatically to new and potentially cheaper and more efficient media. The objective is that the data owner can stop worrying about keeping track of their data and keeping it safe and concentrate on the primary purpose of making effective use of their data.

The system is based around a StorageTek ACS library capable of handling IBM 3480 and 3490 media as well as STK D3-format cartridges. A SUN fileserver running VERITAS software provides access to the system and manages the automatic migration of data between disk and tape. Access is available via high speed network connections into SuperJANET as well as being available via ULCC's network dial up and ISDN services.

For archival storage with virtually no access, this would be a more expensive option. There would be a set up cost of £1500 - £2000 per customer followed by a basic storage cost of approximately £3 – 4 per Gbyte per year. This would cover the cost and replacement of media at regular intervals, periodic integrity checks and refreshing of data and the staff costs of system personnel to operate the system.

A data centre managing a total of 200 Gbytes of data sets would send 2 copies of all media to ULCC. The yearly storage costs including the set up costs would be approximately £3,000 for the first year and then approximately £2,000 per year. A centre with a large collection of digital images comprising 36 Tbytes (see 5.5 above) would be storing 72 Tbytes. The yearly storage costs, including the set up costs, would be approximately £220,000.

For active near-line storage - where the data appears online to the user - a variety of higher price bands are available with subscription levels to the service starting at £1,000 per year and pricing being based on a charge per Gbyte ranging from £40 - £120 per year. Factors affecting the price would include the frequency of access to the data, the speed of access desired and the need for special applications to modify the data before storage or transmission.

ULCC offer two basic styles of service for access to the data repository:

- "bitfile server" – direct access to the virtual filestore offered via any of a number of network protocols. This allows the system to act as a simple repository for data whose content it does not try to interpret. The user decides on filenames and uses them to retrieve files from them later.

- "structured access" – here an application is interposed between the file system and the user. The application has some awareness of the contents of the data and typically hides the file system layout itself from the user who may not need to know explicit file names. Within the application the user can issue queries that relate to the content of the data itself. ULCC consider this style of access may be suitable for data such as an image library or a census database.

Clearly the full ULCC service would be more appropriate for the active management of digital resources for centres rather than purely for archival preservation. ULCC are acting as a distribution centre for the British Library's "Electronic Beowulf" project and are acting as a fully fledged archive for the Public Record Office's CRDA (Computer Readable Digital Archive) project.

If the data centre managing a total of 200 Gbytes of data sets opted for the active storage option from ULCC they would still have to send 2 copies of all media to ULCC. The yearly storage and retrieval costs including the set up costs would be approximately £9,000 – 25,000 depending on the level of access made etc.

6 Summary, Conclusions & Recommendations

6.1 Summary

The terms of reference defined three aims for this study:

- to draw up a matrix of data types and categories of digital resources
- to draw up a decision model for assessing the agreed categories of digital resources to determine the most appropriate method of long term preservation
- to draw up a cost model for comparing the costs of the preferred methods of preservation for each category of digital resource.

In chapter 2 the consultancy defined the scope of the study.

> Section 2.1 references a framework for digital collection management, which comprises seven modules, of which preservation is one.
>
> Section 2.2 defines the key tasks, which comprise digital preservation, and distinguishes the two main approaches adopted to digital preservation.
>
> Section 2.3 defines the three digital preservation strategies which the consultancy was tasked with reviewing.
>
> Section 2.4 describes the approach taken by the consultancy to defining the categories of digital resources covered by this study.

In chapter 3 the consultancy detailed the data types and categories of digital resources covered by the study and the key technical factors which influence how they should be preserved.

> Table 3.1 describes the six basic data types covered by this study.
>
> Table 3.2 describes the ten categories of digital resources covered by this study.
>
> Table 3.3 describes 19 types of application programs used to create the ten categories of digital resources covered by this study.
>
> Table 3.4.1 describes how each of the six data types can be structured for interchange and preservation.
>
> Table 3.4.2 describes how each of the ten categories of digital resources can be structured for interchange and preservation.
>
> Table 3.5 describes how each of the categories of digital resources could have been managed and/or distributed prior to deposit and the influence this would have on how they could be preserved.

In chapter 4 the consultancy describes a decision model which can be used to assess the ten

categories of digital resources and determine the most appropriate method of long term preservation for each. The decision model involves evaluating each category using seven key criteria.

In section 4.1 the consultancy describes the model in detail and explains how each of the seven criteria should be applied.

In section 4.2 the consultancy explains that in the time available the model could only be applied at a very high level for each of the ten categories.

In sections 4.3 – 4.12 the consultancy applies the decision model at a high level to all ten of the categories of digital resources covered by the study.

In section 4.13 the consultancy summarises the results gained in this first attempt to apply the model. For 9 out of 10 categories the preferred preservation approach is "Digital information migration". However, the analysis indicates that this strategy should be implemented in different ways for each of those nine categories.

In chapter 5 the consultancy describes a cost model, which can be used to assess the ten categories of digital resources and compare the costs of the preferred methods of preservation for each category. The cost model references the framework (2.1) that divides digital collection management into seven modules, one of which is preservation. The consultancy reviews each module and assigns costs to each one to arrive at the total costs involved in managing digital collections. The consultancy then reviews these costs and determines whether they are directly or indirectly related to preservation. Those costs that are directly or indirectly related to preservation are totalled up to enable the consultancy to define preservation costs as a percentage of the total costs of digital collection management.

In section 5.1 the consultancy describes some of the issues and resource constraints that impacted the study and restricted the consultancy to applying the cost model to four out of the ten categories of digital resources covered by the study. The consultancy also describes the cost model in detail and explains how preservation costs can be assigned to each of the seven modules.

In sections 5.2 to 5.5 the consultancy applies the cost model to four categories of digital resources covered by the study.

In section 5.6 the consultancy reviews the cost of using third party organisations for off-site archiving and nearline storage of active digital resources.

6.2 Conclusions

The potential audience for this report is a wide one, covering all the following:

- data centres funded by the research councils;
- data centres in the arts and humanities funded by the JISC;
- the Public Record Office and all public and private electronic record centres;
- libraries, museums and art galleries building up digital collections;
- many other organisations building up large collections of digital resources for one purpose or another.

The consultancy believes that the framework defined by Greenstein and referenced in section (2.1) covers all these different digital collection managers. The seven modules should be applicable to all of them.

6.2.1 Defining The Categories of Digital Resources

Given the breadth and range represented by the audience, it is difficult if not impossible to try and define unilaterally all the categories of digital resources managed by these collections and even all the basic data types contained within them. It is certainly impossible to define all the application programs that might have been used to create and structure all the digital resources managed by all these collections. It is also impossible to try and define all the software and systems that may have been used to manage or distribute these digital resources prior to them being deposited at these collections.

However, at the detail level, the collection managers do need to know what basic data types are contained in a digital resource, what applications were used to create it and what systems have been used to manage and/or distribute it. All of these can have a significant influence on how each digital resource can and should be managed and preserved and, most significantly, on how much it will cost to manage and preserve each one.

The consultancy therefore expects that the initial list of ten categories drawn up for this study will be modified and expanded in the light of further research. The aim should be to build on them to create an agreed list of categories that can be referred to and used by collection managers in future.

Similarly the consultancy expects that the list of application program types and management software and systems will be modified and expanded in the light of further research. The aim should be to build on them to create agreed lists that can be referred to and used by collection managers in future.

6.2.2 Defining a Generic Decision Model

Given that the framework (2.1) is generally applicable, the consultancy believes that the decision model defined in chapter four should also be generally applicable to all collection managers.

The consultancy expects that the number and range of categories of digital resources will be altered in future as a result of additional research. However, the consultancy believes that it should still be practical for collection managers to apply the seven criteria in the decision model to any new category of digital resource and use them to decide on the preferred preservation strategy for that new category of digital resource.

Over time it may prove useful to add an extra criterion or divide one major criterion into some subsets but the overall approach behind the decision model should remain applicable.

If, as a result of technological developments or procedural changes, new preservation strategies or new subsets of existing strategies emerge then the work done in chapter 4 applying the model to each of the categories could be repeated, taking into account the new options available to the collection managers.

In the time available, the model could only be applied at a high level by the consultancy. The results achieved and the recommendations made are, therefore, very high level and prone to be too simplistic.

6.2.3 Defining a Generic Cost Model

Given that the framework (2.1) is generally applicable, the consultancy believes that the cost model defined in chapter five should also be generally applicable to all collection managers.

The consultancy expects that the number and range of categories of digital resources will be altered in future as a result of additional research. However, the consultancy believes that it should still be practical for collection managers to apply the cost model to assess any new categories of digital resources and compare the costs of the preferred methods of preservation for each category.

The cost model references the framework (2.1) that divides digital collection management into seven modules. If in future those modules are added to or refined then the cost model would need to be similarly modified but the overall approach should still prove valid.

Clearly the way in which the consultancy has reviewed all the costs associated with all seven modules and determined which costs are directly or indirectly related to preservation can be challenged. It is purely subjective at this stage based on the definition of preservation provided in section (2.2) of the study. Again, however, if collection managers wish to widen or narrow the definition of preservation in future then as long as the new definition is documented and agreed the cost model can still be applied. It simply means that more or less of the costs identified will be allocated to preservation.

If new preservation strategies or new subsets of existing strategies emerge and become the preferred approach for existing or new categories of digital resources, then the work done in chapter 5 applying the cost model to specific categories could be repeated taking into account the new preferred preservation approaches.

In the time available, the cost model could only be applied at a high level by the consultancy to four categories of digital resource. The results achieved and the recommendations made are, therefore, very high level and prone to be too simplistic.

6.3 Recommendations

The British Library Research and Innovation Centre and the National Preservation Office's digital archiving working group should review the study and discuss the findings with Cimtech Ltd.

The study should be published and distributed widely for consultation and comment.

If a consensus emerges in favour of changing the lists of categories, applications or management systems then the suggested changes should be gathered together by BLR&IC.

If a consensus emerges in favour of changing the decision model in any way then the suggested changes should again be gathered together by BLR&IC.

If a consensus emerges in favour of changing the cost model in any way or in changing the way the consultancy has decided whether costs are directly or indirectly related to preservation then the suggested changes should again be gathered together by BLR&IC.

All the changes for which a consensus exists should be passed on from BLR&IC to Cimtech and a short follow up project should be commissioned to edit and republish this study taking into account any commonly agreed changes to the lists, the decision model or the cost model.

If the decision model is generally accepted then a cross section of data centres/record centres/digital libraries should be tasked with applying the decision model to the specific categories of digital resource that they manage. The results should be cumulated and edited and published as a follow up to this study.

If the cost model is generally accepted then a cross section of data centres/record centres/digital libraries should be tasked with applying the cost model to the specific categories of digital resource that they manage. The results should be actual cost figures based on practical experience and should be published as a follow up to this study.

A group of collection managers should be formed comprising at least two specialists with expertise in preserving each of the categories of digital resources that are finally agreed upon. The list would be based on the ten categories listed by Cimtech plus any generally agreed changes – additions or omissions.

The membership of the group should be at least European in scope and ideally international to ensure that it contains the leading experts in each of the categories of digital resource.

At present the group would comprise 20 experts – 2 for each of the 10 categories. The final list may well be significantly longer as additional categories are agreed.

This group should be funded to produce more detailed guides to good practice for the preservation of each category of digital resource.

Finally, the consultancy concluded that there will be a minor role of "last resort" for "technology emulation" and potentially, also for "technology preservation". The consultancy recommended that, for most collection managers who need to rely on these strategies for specific digital resources, the best approach would be to identify third parties that have preserved or are able to emulate the technical environment which they need to run their obsolete application programs.

This is an area where more research is needed, firstly into the feasibility of the "technology preservation" and "technology emulation" approaches as anything more than short term strategies and, secondly, to identify any centres of expertise that can currently provide such services on a commercial basis.

Once the research has been conducted and any existing third party services have been identified, they should be visited and their services assessed by qualified technical experts. At that point an informed decision can be made by the JISC and the National Preservation Office on which of the following policies should be adopted for "technology preservation" and "technology emulation":

- To draw up a list of reliable existing third party services which could be used by collection managers;

- To treat this as an area where further funded research is needed to build up a centre of expertise in the UK that could provide such services to the range of digital collection managers defined above;

- To conclude that both "technology preservation" and "technology emulation" are currently only practical in the short term.

Appendix 1 - References

1 Preserving Digital Information. Final Report and Recommendations. Task Force on Digital Archiving, Commission on Preservation and Access and the Research Libraries Group. Waters, D and Garrett, J (1996)

2 Long Term Preservation Of Electronic Materials: a JISC/BL Workshop Organised by UKOLN. 27-28/11/95. Marc Fresko Consultancy. BLR&D Report 6238.

3 Managing Digital Collections: Towards A Strategic Framework For The Development Of Appropriate & Effective Operational data Policies; Parts 1 and 2. Dan Greenstein, Director, Arts & Humanities Data Services Executive. Arts and Humanities Data Service Executive, Kings College London. Strand London WC2R 2LS UK. . http://ahds.ac.uk/

4 Ensuring the Longevity Of Digital Documents. Jeff Rothenberg. Scientific American January 1995 p 42 – 47.

5 Document management directory. A comprehensive directory of document management and imaging products and services. 9th edition 1998. Roger Broadhurst & Tony Hendley. Cimtech Ltd University of Hertfordshire, 45 Grosvenor Road, St Albans, Hertfordshire AL1 3AW ISBN 0 900458 917. Tel 01727 813651; Fax 01727 813649.

6 Guidelines On The Management Of Electronic Records. Public Record Office EROS Programme. Consultation Draft. Ian Macfarlane, Public Record Office, Kew, Richmond, Surrey TW9 4DU. Tel 0181 876 3444.

7 Guide to Depositing Data. The Acquisition Section, The Data Archive, University of Essex. Wivenhoe Park, Colchester, Essex. CO4 3SQ. http://hds.essex.ac.uk/

8 Image Scanning: A Basic Helpsheet. Electronic Text Center, Alderman Library, University of Virginia, Charlottsville VA 22903 Tel 804 924 3230. .

9 Special Collections. Digital Image Creation. David Seaman. Electronic Text Center, Alderman Library, University of Virginia, Charlottsville VA 22903 Tel 804 924 3230

10 Image and Multimedia Database Resources. Howard Besser and Rachael Onuf.

11 Framing the Picture: standards for imaging systems. Jennifer Trant. Paper presented for the International Council on Hypermedia and Interactivity in Museums. Museum Computer Network Joint Conference. San Diego California October 1995.

12 Benchmarking image quality: from conversion to presentation. Anne R Kenney. Associate Director, Department of Preservation & Conservation, Cornell University.

13 Leave it to the labs? Options for the future of map and spatial data collections. C R Perkins. The LIBER Quarterly Vol 5 1995 No 3 p 328. Dr H Schnelling, Universitatsbibliothek GieBen D-35386 GieBen Germany. Fax +49/641/46406 Email:

14 NERC Data Policy Handbook. Natural Environment Research Council. Version 1.0 January 1996. NERC Data Strategy Group, c/o NERC Scientific Services Polaris House North Star Avenue Swindon SN2 1EU Tel 01793 411683; Fax 01793 411610.

15 Metadata – Library & Information Briefing 75. Rachel Heery, Andy Powell & Michael Day, UKOLN. ISSN 0954 1829 Library Information Technology Centre. TBC Distribution, South Lodge, Gravesend Road, Wrotham, Kent TN15 7JJ. Tel 01732 824700; Fax 01732 823829. Email: tomlinsons@easynet.co.uk

16 Data Archives: Resource centres for teaching and research. Bridget Winstanley, Assistant Director, Data Archive, University of Essex. Library & Information Briefings. Issue 52 April 1994. ISSN 0954 1829. Library Information Technology Centre, TBC Distribution, South Lodge, Gravesend Road, Wrotham, Kent TN15 7JJ. Tel 01732 824700; Fax 01732 823829. Email: tomlinsons@easynet.co.uk

17 Maintenance and preservation of large databases. Denise Lievesley, UK Data Archive. The LIBER Quarterly Vol 6 1996 No 4 p 472 – 482. Dr H Schnelling, Universitatsbibliothek GieBen D-35386 GieBen Germany. Fax +49/641/46406 Email:

18 Data Management For The OMEX 1 Project: A Case Study. Roy K Lowry. British Oceanographic Data Centre, Proudman Oceanographic Laboratory, Bidston Observatory, Birkenhead, Merseyside L43 7RA UK Tel 0151 653 8633; Fax 0151 652 3950 Email

19 Oxford Text Archive. OTA Collections Policy Version 1. 22/12/97. The Oxford Text Archive. Oxford University Computing Services, 13 Banbury Road, Oxford, OX2 6NN. Tel 01865 273238; Fax 01865 273275;

20 Discovering Online Resources Across the Humanities. A Practical Implementation of the Dublin Core. Edited by Paul Miller and Daniel Greenstein on behalf of AHDS and UKOLN. 1997. ISBN 0 9516856-4-3. The Arts and Humanities Data Service, Kings College London, Strand, London WC2R 2LS. 95 pages. .

21 NARA Guidelines for Digitizing Archive Materials for Electronic Access. Stephen Puglia, Preservation & Imaging Specialist & Barry Roginski, Computer Specialist. National Archives and Records Administration, 8601 Adelphi Road, College Park, MD 20740, USA.

22 Digital Formats For Content Reproductions. Carl Fleischhauer, Technical Coordinator, National Digital Library Program, Library of Congress. August 1996.

23 The Setup Phase of Project Open Book. Paul Conway and Shari Weaver, Yale University Library. June 1994

24 Yale University Library's Project Open Book. Preliminary Research Findings. Paul Conway, Head, Preservation Department, Yale University Library. D-Lib Magazine. Feb 1996. ISSN 1082 9873. Paul.
http://

25 Visual Arts Data Service. Draft Guidelines for Depositors. Catherine Grout. October 1997.

26 Access Aids and Interoperability. Caroline R.Arms. National Digital Library Program, Library of Congress. August 1997.

Appendix 2 - Bibliography

1 Preservation Management. Graham Matthews, Dept of Information & Library Studies, Loughborough University. Library & Information Briefings Issue 73. August 1997 Library Information Technology Centre. ISSN 0954 1829. 18 pages TBC Distribution, South Lodge, Gravesend Road, Wrotham, Kent TN15 7JJ. Tel 01732 824700; Fax 01732 823829. Email: tomlinsons@easynet.co.uk

2 National Library of Australia. National Preservation Office 1997. Statement of principles. Preservation of and long term access to Australian digital objects. <http:// >

3 Preservation and the management of library collections.
Feather. J 1996. 2nd ed Library Association Publishing.

4 Matthews, G, Poulter, A and Blagg, E (1997) Preservation of digital materials. Policy and strategy issues for the UK. London, British Library Research and Innovation Centre. JISC/NPO Studies on the Preservation of Electronic Materials. BL Research & Innovation Report 41.

5 Introduction to Imaging; Issues in constructing an image database
Howard Besser and Jennifer Trant

6 Archaeology Data Service – Guidelines for Depositors v 1.0. Archaeology Data Service, University of York, King's Manor, York YO1 2EP Tel 01904 433954; Fax 01904 433939. Email

7 Archaeology Data Service – Guidelines for Cataloguing Datasets with the ADS v 1.0. Archaeology Data Service, University of York, King's Manor, York YO1 2EP Tel 01904 433954; Fax 01904 433939. Email

8 Archaeology Data Service – ADS Collections Policy v 2.0. Archaeology Data Service, University of York, King's Manor, York YO1 2EP Tel 01904 433954; Fax 01904 433939. Email

9 Resource Discovery Within The Performing Arts. Malcolm Jones. Performing Arts Data Service. University of Glasgow. Glasgow G12 8QQ. Tel 0141 330 4357; Fax 0141 330 3659.

10 A Framework Of Data Types & Formats & Issues Affecting The Long Term Preservation Of Digital Material. John C Bennett. BLR&IC Report 50. 1997.

11 Automated SGML Tagging Standard Format For hardcopy US Patent Applications Distributed Object Computation Testbed (DOCT) Technical Report. Task Group A – Develop document and data models. Prepared by SAIC (Science Applications International Corporation).

12 Concept of Operations for the Distributed Object Computation Testbed (DOCT)

Project of DARPA (Defense Advanced Research Projects Agency) and US Patent & Trademark Office.

13 Guide For Managing Electronic Records From An Archival Perspective. ICA Committee On Electronic Records. Feb 97. ICA Studies CIA 8.

14 Digital Historical Collections: Types, Elements & Construction. Carl Fleischauer. Library of Congress. August 1996.

15 The Digital Preservation Consortium Mission and Goals. (now the National Digital Library Federation). Council on Library and Information Resources (CLIR), 1775 Massachusetts Avenue N.W.Suite 500, Washington D.C. 20036-2188. Tel (202) 939 4750. **Error! Reference source not found. Error! Reference source not found.**

16 Preservation in the digital world. Paul Conway. Yale University Library. Publication 62, CLIR (see above). **Error! Reference source not found.**

17 Digital image collections. Michael Esther. Publication 67, CLIR (see above). **Error! Reference source not found.**

18 Digitization as a means of preservation? Hartmut Weber and Marianne Dorr, Publication 69, CLIR (see above). **Error! Reference source not found.**

19 Intellectual Preservation: Electronic Preservation Of The Third Kind. Peter S Graham. Rutgers University 1994. **Error! Reference source not found.**

20 E- Recs Final Report Committee on the Records of Government. Daniel Atkins & Margaret Hedstrom, , University of Michigan. **Error! Reference source not found.**

21 Final Report Of Cimtech Study On Preservation Of Digital Material For British Library Nov 95. (internal report).

22 Berkeley digital library's collection policy. **Error! Reference source not found.**

23 ICPSR's Policy on Data Media and Distribution. Inter-university Consortium for Political and Social Research. 1997. **Error! Reference source not found.**

24 Preparing Data For Archiving: What ICPSR Requires. **Error! Reference source not found.**

25 ICPSR Guide to social science data preparation and archiving. 1997. **Error! Reference source not found.**

26 Project Open Book World Wide Web Home Page **Error! Reference source not found.**

27 Image Formats For Preservation and Access Michaael Lesk. Commission for Preservation and Access. TAAC. 1990. **Error! Reference source not found.**

28 Ackerman, Mark S and Fielding, Roy T. Collection maintenance in the Digital Library, Proceedings of Digital Libraries 95. Austin TX June 1995 p 39-48. **Error! Reference source not found.**

29 Kenney, Anne R. Digital to microfilm conversion: an interim preservation solution. Library Resources & Technical Services 37 (1993) 380-401.

30 Robinson, Peter. The digitisation of primary textual sources. Office for Humanities Communication Publication No 4. Oxford. Oxford University Computing Services 1993.

31 Van Bogart, John W. Magnetic tape storage and handling: a guide for libraries and archives. Washington D C CLIR. CPA 1995. **Error! Reference source not found.**/

32 AHDS Guides To Good Practice **Error! Reference source not found.**

33 Australian National Library Selection Committee on On Line Australian Publications – Guidelines for the Selection of Online Australian Publications Intended for Preservation by the National Library. http://**Error! Reference source not found.**

34 The Electronic Text Center Guide to Document Markup and Preparation 1994 – Present. David Seaman. **Error! Reference source not found.**

35 Margaret Hedstrom. Digital preservation: a time bomb for Digital Libraries. **Error! Reference source not found.**

36 Janice Mohlhenrich ed Preservation of Electronic Formats: Electronic Formats for Preservation. Fort Atkinson WI Highsmith Press 1993

37 Ann Gerken Green & Jo Ann Dionne. Preserving the Whole: A two track approach to rescuing data and metadata. Interim Report to the Commission on Preservation and Access Dec 1996. **Error! Reference source not found.**

38 AHDS list of major initiatives and information resources in digital preservation **Error! Reference source not found.**

39 Digital Libraries Initiative. NSF; Dept of Defense Advanced Research Projects Agency (DARPA) and NASA in US **Error! Reference source not found.**

40 National Digital Libraries Federation 15 US research libraries – "agreed to cooperate on defining what must be done to bring together digitised materials that will be made accessible to students, scholars and citizens everywhere and that document the building and dynamics of US heritage and cultures. **Error! Reference source not found.**

41 Clifford Lynch & Hector Garcia-Molina. Interoperability, Scaling and the Digital Libraries Research Agenda **Error! Reference source not found.**

42 Delivering Technology for Digital Libraries: Experiences as Vendors. William T Crocca and William L Anderson. **Error! Reference source not found.**

43 The Digital Research Library: Tasks & Commitments. Peter S Graham. **Error! Reference source not found.**

44 Public access to the government's electronic documents (EROS-IPC) Kenneth Tombs Independent Consultant. IMAT Vol 29 no 5 1996 p 193 – 195. ISSN 0266 6960

Cimtech Ltd, University of Hertfordshire, 45 Grosvenor Road, St Albans, Hertfordshire AL1 3AW. C.Cimtech@herts.ac.uk

45 Preservation of digital material for libraries. John Mackenzie Owen. The LIBER Quarterly Vol 6 1996 No 4 p 435 – 451. Dr H Schnelling, Universitatsbibliothek GieBen D-35386 GieBen Germany. Fax +49/641/46406 Email: **Error! Reference source not found.**

46 Digitisation and preservation in the French National Library. Daniel Renoult. The LIBER Quarterly Vol 6 1996 No 4 p 465 – 471. Dr H Schnelling, Universitatsbibliothek GieBen D-35386 GieBen Germany. Fax +49/641/46406 Email: **Error! Reference source not found.**

47 Preservation and digitization; principles, practice and policies. The 1996 National Preservation Office Conference. Paper Preservation News No 80 December 1996 p 13 – 14.

48 Towards a universal data format for the preservation of media. Dave MacCarn SMPTE Journal July 1997. P 477 – 479

49 JSTOR: the Andrew W Mellon Foundation's Journal Storage Project. R D Gennaro. Proceedings of the 19th International Essen Symposium 23-26 September 1996. Ed by Ahmed H Helal & Joachim W Weiss. Essen Germany. Universitatsbibliothek Essen 1997 p 223 – 230.

50 JSTOR: Building an Internet accessible digital archive of retrospective journals. R D Gennaro, 63rd IFLA General Conference Proceedings Aug 31 – Sep 5 1997

Appendix 3. Table of digital preservation cost elements

Compiled by Neil Beagrie, Daniel Greenstein, and the Arts and Humanities Data Service

The following table sets out the cost elements involved in developing and preserving digital collections and shows contingencies which may affect the actual cost placed against any one of the elements. It is based closely on the stages involved in the life-cycle of a digital resource from its initial design through to its long-term disposition. Any data managing agency which uses the table to calculate preservation costs will require collections policies which indicate the kinds of data resources are suitable for inclusion within their collections, and clearly articulated recommendations with regard to standards and best practices either required or preferred from data depositors/suppliers. In all cases actual charges will vary depending upon whether the data resource in question conforms to preferred or required standards and best practices, and to the contingencies identified in the table below. Consequently, any value entered against a cost element is likely to appear as a range rather than as a single numeric value.

Costs involved in digital preservation

Life Cycle Stages	**Cost Elements**	**Description/purpose of eleme**
1. Data design, data creation		Encourage adoption of service-preferred standards and thus reduc long-term preservation costs
	1. Publications	Targeting data creators / suppliers and informing them about preferre standards and best practices
	2. Training events	Ditto
	3. Consultancy	Direct involvement in an advisory capacity within a data creation project.
2. Data accessioned into collections		
	1. Acquisition	i. Identification of potential acquisitions
		ii. Negotiating with potential depositors
		iii. Evaluating resources for accession
	2. Accessioning	i. Verifying data against data documentation; reading, copying, and validating files

		ii. Checking inventories, schedules and licences supplied; administeri deposit and if necessary checking with depositor over errors and omissions
	3. Catalogue records and documentation	Creating or enhancing records and documentation
	4. Data Processing	Technical manipulation or conversion of data deposited/supplied
	5. Data Storage	Storing data in both preservation a distribution formats (where different), offline/offsite copies, a ensuring viability of storage media
	6. Preservation	Cost of preservation strategy (strategies) viz migration, emulati technology preservation, preservat of data at bit level, etc.
	7. Monitoring Reports	Reporting on usage statistics to depositor
	8. Interface design	Required only for data intended fo online distribution
	9. Administering commercial use	Acting as broker between depositor/supplier and users requesting commercial use (includ charging and payment of royalties etc)
	10. Withdrawal fees	Charge for value added to data through archival process, management and dissemination
3. Data Use and Administration		
	1. Distributing data and documentation	Delivery of data to users online or other means
	2. Access and administration	Administration of orders, registration, licensing, and invoici Provision of staff and technical infrastructure to access collections
	3. User support	Staff research and searches, data

		subsetting of collections, lost passwords, certification of digital copies for legal use, technical and subject specialist advice
	4. Royalties	Collecting and distributing these where required by third party suppliers
	5. Training	Targeting users and encouraging their use of the resources
	6. Publications	Ditto