British Library Research and Innovation Report **50**

JISC/NPO STUDIES ON THE PRESERVATION OF ELECTRONIC MATERIALS

A FRAMEWORK OF DATA TYPES AND FORMATS, AND ISSUES AFFECTING THE LONG TERM PRESERVATION OF DIGITAL MATERIAL

John C Bennett

British Library Research and Innovation Centre
1997

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 Archive 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/304

ISBN 0 7123 3312 6

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

Table of Contents

Figures and Tables

Working Papers

Executive Summary

The aim of the study is to develop a framework which can help manage the resolution of the issues associated with the long-term preservation of digital material. Although a great deal has been discussed and written about digital material preservation, there would appear to be no overall structure which brings together the findings of the numerous contributors to the debate, and allows them to be compared. This Report attempts to provide such a structure, whereby it should be possible to identify the essential elements of the preservation debate and to determine objectively the criticality of the other unresolved issues. This Report attempts to identify the most critical issues and employ them in order to determine their affect on preservation practice. Where possible, the management issues and recommended approaches are high-lighted where they occur. For the purposes of clarity, some of the issues are documented as two working papers attached to the report.

In the report the phrase "long term" is equated to 50 years, as a working hypothesis, a period of time which takes us towards the limits of one of today's most durable storage media: compact disk storage technology. 50 years ago the first commercial digital computers were under development.

The diagram (Figure 1) represents the three aspects of the Framework.

a) the use of a two-by-two matrix in order to review the provenance of the item and the conditions that relate to its current and future use (held as Working Paper 1)
b) a process to evaluate the characteristics of the item, in order to determine both its sensitivity to technological obsolescence and its inherent need for specialist attention prior to preservation (the main body of the Report), and
c) issues relating to the governance of the archived item, and the requirements of the item during its life-cycle within the archive (held as Working Paper 2).

At the outset of the study, the team predicted a matrix of least three dimensions, similar to a decision table that could be used to determine the approach taken to preservation for candidate items of digital material. After further examination of a wide variety of digital material, and based on past experience, the team concludes that in order to achieve cost-effective long term preservation (achieving permanence) it is essential to "Keep Things Simple, Sir!". The KISS principle is not a new concept in computer management circles, yet remains highly effective in information engineering management. From the perspective of cost management alone, the KISS principle predicates that digital archive material should be held in archive in a standard format, on standard media, and managed by one of a few standard operating systems. Material that does not conform must either be processed prior to preservation or be managed under a different regime, with a premium scale of charges.

The study concludes that the overall management task in long term preservation is to moderate the pressure to preserve (Step 1) with the constraints dictated by a cost-effective archive (Step 3). This continuing process of moderation is documented through the Scorecard (Step 2 - the subject of this Report).

Initiating and maintaining the second step is therefore most critical to the practical application of the whole Framework. A set of matrices calculates the complexity of an object's preservation needs, based on its characteristics. The resulting Preservation Complexity Scorecard for an object helps identify the preservation approach, and special cases are identified during the scoring process. Over time the Scorecard calibration will change as new digital technologies are used to access and preserve digital material. The ideal archive environment will also change over time.

As an adjunct to the process of Technology Watch, the Scorecard can act both as a trigger to rescue items that are in danger of being lost through technological obsolescence, and as a note to point up opportunities to move to cheaper, more durable storage media. The Scorecard is intended to be a part of a living document, published widely and annotated as special needs are identified in particular circumstances. Overall its aim is to reduce the complexity of the preservation environment, by identifying the commonality of preservation issues, and by helping to initiate solutions and corrective actions.

In summary, by assessing the digital material's provenance and conditions of use, the Scorecard determines the governance and requirements of the archive. By knowing more about the candidate digital material, the probability of a successful archive and retrieval in the long term are greatly enhanced. By improving the management of the archive, the Scorecard can be simplified and more digital material can be archived more securely with the same resources.

The Study overall recommends that a work programme should be started to:

a) Establish a Scorecard approach (to measure preservation complexity),
b) Establish an inventory of archive items (with complexity ratings) and
c) Establish a Technology Watch (to monitor shifts in technology), in order to be able to manage technological change

and in support of this,

a) establish a programme of work to explore the interaction of stakeholders and a four level contextual mode in the preservation process

Section 1. Introduction

Terms of Reference

The study is part of a wider programme of studies, funded by the Joint Information Systems Committee ("JISC"). The programme was initiated as a consequence of a two day workshop at Warwick University, in late November 1995. The workshop addressed the Long Term Preservation of Electronic Materials. The attendees represented an important cross-section of academic, librarian, curatorial, managerial and technological interests. 18 potential action points emerged, and these were seen as a basis for initiating further activity. After consultation, JISC agreed to fund a programme of studies.

The aim of the consultancy work for this Study is to

a) devise a topology, or framework, of the data types and formats within the digital domain
b) indicate the likely problems, requirements, issues and responsibilities appropriate to each category
c) identify the most appropriate method of preservation for each category of digital material
d) propose the most appropriate method of managing the process in the interest of the stakeholders

Method of Working

Background reading, both on paper and through the InterNet, provided a wide range of source material. Discussing the issues with the study's consultative committee and practitioners raised another range of practical issues not always reflected in the conference papers. Brainstorming within the team followed, attempting to bring some structure to the quantity of information that had been gathered. Compiling the Report took far longer than planned, owing in part to the scope of the study and the need to do justice to such a widely debated topic.

Deliverables

A framework was envisaged from the outset as the best way of representing the different types of material and how they should be handled from the aspect of preservation. The complexities that emerged during research, suggested that a fixed set of matrices (or decision table) could not represent the full scale of interactions between the various components of the study. The study Report therefore grew in breadth to cover the management of the process, providing a framework not just for the material's formats and preservation requirements but also a framework for the management of the archive and the discussion of the principles implied.

Acknowledgements

After several days' worth of trawls on the InterNet, the team felt it was going round in circles, the same names kept appearing on different search engines, intriguing papers were sometimes on inaccessible Web sites, and sometimes the team faced server time-outs just as the interesting paper was being retrieved.

The study's consultative committee of Chris Rusbridge, Nancy Elkington, Dan Greenstein and Michael Alexander were very helpful in initiating the information search. They started the debate for the team on the importance of Intellectual Context, post hoc rescue, and the impermanence of digital material. If the study could have been longer, it would have been very valuable to have been in permanent discussion with them!

The major conference and meeting papers that helped set the investigative process going and stimulated the team to strongly agree or disagree with current thinking were:

The JISC / British Library Workshop at Warwick University, November 1995
The Report of the Taskforce on Archiving of Digital information, May 1996
Policy and Strategy Issues for the UK, follow-up meeting December 1996

From the 10 cm of paper (printed out) from the Web, the team is indebted to the following authors for their particular insights, which the team used throughout this Report:

David Bearman, Item Level Control and Electronic Record-keeping, August 1996
Dr. Cameron Easton, Principles of Preservation, September 1992
Maggie Exon, Long Term Management Issues in the Preservation of Electronic Information, November 1995
Peter Graham, Long Term Intellectual Preservation, March 1994

Next Steps

The study's consultative committee will discuss the report and its issues and recommendations that the study Report in order to determine how to take forward the rest of the JISC study programme.

Section 2. Overall Context for Preservation Issues

The need to manage the preservation of digital material both immediately and in the long term has encouraged the promotion of a wide range of approaches and the proliferation of a diversity of discussion topics. The debate of the critical issues has been overwhelmed by organisations protecting their turf and collection managers jostling for project financing.

Posing four main questions may help provide an overall context for the discussion and resolution of preservation issues that are connected with digital material (Figure 2). For this Study, the key question is Question 4 - Where should we keep our archived material? The other questions provide a context both to examine the issues, and manage the related JISC-funded studies (explored further in Section 11).

Question 1: Why?

Preservation is a response to the threat of destruction. Some individual ultimately must initiate the response when the threat has been recognised, and the scale of the reaction may be in proportion to the value that is placed on the object under threat. Their reaction incurs a cost which will continue to be incurred, while the threat appears to remain. Funds will be drawn upon, and resources will be mobilised, that have been held in reserve explicitly for the purpose of preservation. Other stakeholders are drawn into the preservation activity as time passes.

In the world of digital material, the old rules do not apply so clearly. The cost and effort required to preserve candidate digital material may not be proportionate to the value of the material, nor are they related directly to the urgency created by the threat. Whereas before benign neglect of printed paper-based material was a viable course of action, and a delayed reaction could in itself be an act of preservation. With digital material, decisions are required, supported by authorised expenditure, to enable resources to be deployed quickly in order to counter the threat of irreversible loss. The resources may involve substantial capital investment as well as specialist labour, both available in the near term only at a premium.

The contest for limited resources and the balancing of conflicting priorities translates into a question of selection: “why should this digital material be preserved?” The solution is no more straightforward for any collection developer, though with digital material, the threat of loss and the volume of material requiring attention is growing year by year. From this initial “why?” other questions grow, questioning the long term viability of any stored information, and the cost and benefits of preservation action.

a) What is the rationale for preservation?
b) When an object is retrieved from the archive will it still be valuable in 50 years time? Will it still be recognisable and comprehensible?

c) Research libraries and Legal Deposit libraries have very different requirements when retaining material over long periods of time. In each case, what costs are non-discretionary, how do they apply to an item's life-cycle in archive, and when will costs start to be discretionary?
d) What benefits are measurable, how can they be achieved, and who can be tasked with capturing them?

Question 2: How much?

Because so much of a digital item is connected to its immediate technical regime, the preservation specialist is concerned not to leave out any information that will later prove to be valuable. Only what is sufficient and necessary should be carried forward. The preserved material, held as though in a sealed capsule, must be accompanied by material that moves forward technologically in step with the changing world, changing its format and style, while still being able to fulfil its purpose. Otherwise when the capsule is opened the instructions on how to use the material may be in a perfect state of repair, but all the same incomprehensible. In other words in order to preserve the integrity of digital material, the surrounding medium may need to be changed frequently, losing data in the process. The question of "how much" leads into the wider debate of the long term marriage of unchanged material with material that must change.

a) What contextual information is necessary for preservation?
b) It is not sufficient to register and index an object, it must carry extra information with it into the archive, what contextual information is sufficient, so that when it is retrieved it can be interpreted correctly?
c) How the object will eventually be accessed, and for what purpose, how will this affect the approach to preservation?
d) While the object may need to be held unchanged, while it is in the archive the media on which it is stored may need to be upgraded every five years. What is the interplay of these two principles?

Question 3: How?

Having determined the contents of the sealed capsule and the accompanying contextual material, there remains the task of capturing and storing the materials in the archive. The straightforward process of managing the archive is complicated by the possibility that opportunities exist for archive managers to avoid using the procedures rigorously. The risks may be negligible if the procedures are established on the basis that human error and mechanical failure are inevitable. Unlike the existing national archives of printed material, the value of the digital material may not be directly proportionate to its age or cost of production. Digital material may need to be treated exactly the same, whatever its provenance. How can these good practices be established?

a) What are the preservation processes' procedural needs in order to achieve a long term archive?

b) Who are the stakeholders, who will influence the way the archive is built up and managed?
c) What quick, cost-saving routes are there, which do not adversely affect the quality of the archive?
d) What safety nets exist which can provide a fall-back for the archive should accidental loss or deliberate sabotage to the archive occur?

Question 4: Where?

All technology consists mainly of electronic storage used for different purposes, having different orders of size, security, and cost. If storage technology is ubiquitous, the question is not when to archive but where is the best place to create a preserve of digital material? Is it a place where little changes over time? or should it be in the centre of the latest networked configuration?

a) While technology is in a state of continuous transition, when will technology be resilient and stable enough for any item to be assured of its long term preservation?

Conclusion

The four questions are not intended to act as a straitjacket on opinions and ideas. They are aimed primarily controlling the scope of this study and co-ordinating the efforts of future, subsequent studies. The questions and their scope are also intended to encourage effective debate, expedite actions and avoid delay in all matters relating to the long term preservation of digital material.

Components of the Framework

Summary of Sections 3 - 7

Taking as a starting point the present day, Question 4, the second step in the Framework, assesses the complexity of the candidate digital material to be preserved by examining the type of material, the type of format, the current media used to hold the material and the platform on which it currently resides. The complexity is registered in each category, in the first two by a score out of five. In all four scores make up the Scorecard for the candidate material. In the analysis, complexity factors of the actual occurrence of the material are noted, when they may materially affect the outcome, by understating or overstating the combined complexity rating.

The problem cases, or high scoring candidates, can be defined as where:

a) the potential for loss is high, through technological obsolescence or the volume of data to be preserved
b) there is an in-built dependency on the surrounding infrastructure, for example, databases in general and GIS databases in particular
c) embedded programs, compression routines, macros and executable code may be hidden, and the code is not transferable across technology boundaries.

Looking to the future, the Scorecard should be used as reference, first to see whether the scoring continues to be accurate, and second to build up a case history for future benchmarking. Use of the Scorecard approach allows the cross-referencing and checking of similar cases over time and across platforms, in order to both track technology shifts and validate the core assumptions. The Scorecard will also be affected, potentially simplified with lower scoring in all categories, by technical advances in the archive environment. The approach is open-ended, allowing for future expansion, as the diversity of candidate digital material increases.

The Scorecard is the repository of the findings of the Technology Watch. It alerts preservationists to trends in technology diversity which will lead to step changes in software functionality, which will lead ultimately to loss of access to archived items. The Scorecard can also alert collection developers to step changes in the management of the archive. In future these may permit the storage of more diverse formats than at present, and could reduce the amount of pre-preservation processing.

In summary, the Scorecards reflect the Principles of the Framework (avoiding obsolescent technology, using enduring file formats, ensuring the long term provenance and value of the data) and will also be modified as necessary by advances in the practice of archive management.

Section 3. Dimension 1: Type of Material

The first major factor affecting the approach to the preservation of digital items is the type of material (Figure 3). For example, textual documents are possibly the simplest items to preserve, they are well-scoped, containing all the information relating to document within the file, when it is presented for preservation. Complexity remains low if they use a standard mark-up language. The complexity rating rises when a document links to other objects, outside itself, or when essential extra functionality for document formatting (Table of Contents) is introduced, or the document contains a macro, or the document is intended to work in a networked environment and contains HTML linkages. The risk is that some of these features may not be reproducible, or inaccurately, in the future. Either way the evidential nature of the record is diminished, potentially catastrophically. At present, when such documents are retrieved, these functions are usually lost, and the unformatted text is displayed or "default" templates are used. The loss is restricted to formatting and presentation.

In the matrix (Figure 3) each type of material is given a base score (1 being the least complex to preserve, and 5 the most complex). To the base score is added a complexity factor, triggered by some functionality feature that adds cost and effort ("difficulty") to the handling of an item when preserving it over the long term.

GIS databases are the most complex, partly because of the inter dependence of the components that make up the final overlaid database and backdrop, but also because there is a multiplicity of standards for mapping access and storage. GIS databases can also be very large. This features cause Image, Sound and Video to be marked up as having the next highest score in complexity.

The development of "Office Suites" has increased the number of cross-object connections, as well as supporting more integrated, encapsulated, holdings of information, equivalent to "bound" volumes. The amount of complexity that such advances introduce is dependant on how they will be supported in future. The linkages are becoming standardised by proliferation and use (de facto), though making them de jure is always a much delayed, prolonged effort, usually producing too little, too late. Therefore, by proliferation, the facilities will become embedded and supported in more and more products, irrespective of vendor. The material and the format become bound together, similar to a book. From the perspective of preservation and future access, the resources that maintain the usefulness of the material have been "donated" by the software vendors, although from self-interest. Technology Watch will monitor their continued willingness to donate the resources without change and without major change.

Material	Base Score	Complexity Factors (add to the base score)	Risk
Text / Document	1	Functionality (+1), Macros (+1), Templates (+1)	Loss of format
(encapsulated)	1	Linkages (standard)	Loss of links
	2	Linkages (+1), HTML (+2)	Loss of external data
Spreadsheets	1	Formatting (+1)	Loss of format Loss of meaning
Multiple Spreadsheets	2	Linkages (+1), Macros (+1)	Loss of external data
"Office Suite" documents	2	Links, Views, Indexes are standard	Loss of access to all items Loss of meaning
Database records	3	Structures and rules (+1)	Loss of meaning
	3	Indexes (+1) Sub-routines, external links (+1)	Do not store, recreate index
Maps (raster)	2	Colour encoding (+1)	Loss of image quality
Maps (vector)	3	Non-standard calculation or base grid (+2)	Ambiguity of plotting
GIS Database	4	Mapping to underlying raster or vector Map	Ambiguity of plotting
Image Sound Video	1	Linkages (+1)	Loss of links
	1	CIP format not yet standard, Packaging has value (+2)	Loss of meaning
	1	Variation of encoding (+1)	Loss of image quality
	1	H/w-based compression routines (+3)	Loss of "key" to decompress
	3	Very large uncompressed size (+2), e.g. X-rays	Specialised archive s/w required
Image database	3	"Fuzzy" Search software	Do not store, recreate

Figure 3 - Scorecard: Types of Material

Section 4. Dimension 2: Type of File Format

The second major factor affecting the approach to the preservation of digital items is the type of file format (Figure 4). Most software developers aim to make their products able to import and export objects into many different formats. The translation from one into the other is not fool-proof but it serves the purpose of exchangeability. With preservation of evidence in mind, the future decoding of these formats will be critical. Exchangeability between current software products is not a sufficient mechanism to provide permanence.

Formats become someone else's problem when the item to be preserved is held within a capsule, such as Microsoft Office or Lotus Notes, from which it can be redisplayed. The translation or display is the responsibility of the enveloping software. The envelope will evolve over time, but the translation techniques will be preserved. Word Version 6 will always be associated with a Word Version 6 "launch" or "view" software module. In the same way, other formats which are standard at the time of capture provide the least risk path for preservation. Documents in a non-current, non-standard formats cannot be stored in their native format, unless a "launch" or "view" mechanism can be stored in a capsule with them. With text-based documents, the rules of evidence do not require the archive to retain the original data with its full format characteristics. It is necessary instead to provide supporting evidence that the text could not have been amended during the time the item was in the archive.

The most complex to guarantee preservation at present are the graphics formats. They are continuously evolving, and the evolution has still some time to run. Each software company adapts standards to suit their product, to limit the problems of upward compatibility, and to enforce customer loyalty. It is too soon to predict that formats, promoted as open and potentially non-proprietary, even valuable contributions to inter-changeability such as PDF, are here to stay as a long term standard. A "launch" or "view" facility will have to be stored with them, or the graphic objects will have to be stored in a non-proprietary format, TIFF or BMP.

As an recent example of the transitory nature of some of these graphics "standards", the popular GIF format, a common element on the CompuServe network, is now hardly used. This is a direct result of CompuServe being forced to stop using it as a result of copyright infringement law suit, successfully brought against them. GIF has now been replaced by a similar, but different format, invented by CompuServe. The net effect is that after a short period of time (possibly measured in terms of use of the network: 10 million on-line messages?) the old GIF will not be supported, under the terms of the settlement. Some stored CompuServe message attachments may therefore become garbled or inaccessible.

The working paper (Working Paper 3) attached to this report demonstrates the variety of formats in the graphics area and the allowable sub-types that exist within them. It is very difficult to place a general format within one category, a suffix, such as TIFF, can have many internal formats, all slightly tuned for a particular software product or environment. The differences only become apparent when an image is being manipulated, compressed or edited.

Format	**Base Score**	**Complexity Factors (add to the base score)**	**Risk**
Recognised uncompressed standard formats	1	Variants on standards are common, but usually do not prevent retrieval (+1)	Loss of quality if lower bits-per-pixel chosen
Recognised standard document-level formats	1	Products known to be rarely used or obsolete (+2)	Loss of data
Recognised Meta and Vector formats	2	Variants on standards are common, but usually do not prevent retrieval (+1)	Loss of data
Recognised compressed graphics formats	2	Products known to be rarely used or obsolete (+2) Products have special compression routines (+2)	Loss of data or translate into portable format
Proprietary-based formats or languages of any of the above	5	Complexity will vary dependant on the routines available to bridge to more standard formats (+ or -), for example, proprietary fractal compression algorithms.	Loss of data and meaning, loss of resolution on output

Figure 4 - Scorecard: Types of File Format

Overall, apart from the effects of the software market's internecine warfare, the main division of formats for the future are lossless formats (whether compressed or uncompressed) and lossy formats. The degree of loss in lossy formats is only of concern to the preservation environment if the uncompressed object cannot serve the purpose for which it was preserved, for example, as evidence, supported by a adequate copy or facsimile of an original. Loss is an issue for lossy formats when:

a) through the passage of time, embedded filenames and locations change or become defunct, having been unrecognised during a previous trawl to manage an update of all known references
b) sudden step changes, such as GIF and CompuServe, which prohibit from a certain date the use of a particular format, and the software has been withdrawn
c) emerging InterNet usage popularises new improved formats, which do not cater for the older formats
d) the greater degree of compression leads to a greater degree of "wobbliness" during processing the image.

Section 5. Dimension 3: Type of Media

In the last 50 years, the diversity of media on which data has been stored has not diminished, but increased. Despite the diversity, the most durable of media remains the tape. The original tape storage mechanisms have changed size, recording density, encoding, capacity, speed and reliability, so that they are no longer recognisable. After tape, the disk is the most durable, but the disk has changed more radically than the tape. Optical, magneto-optical, magnetic and solid state devices now compete to hold commercial data. Many other variants and hybrids of these two dominant technologies and others have missed becoming museum pieces, but their legacy in terms of data storage remains.

In Figure 5, there is no base score, because it is the opinion of the team that there is no real choice over the ideal media for long term preservation of digital material. The media that should be used is either 8mm DAT volumes or some derivative of CD (a new CD format may require bulk copying of data). The two technologies combine portability, reliability, speed of access and a greater capacity. They score highly in the "Capture" and "Storage" categories of the archive (Figure 6), because of their longevity, portability and lack of susceptibility to damage. All other devices do not score as highly.

Media	**Example**	**Complexity Factors (no score)**	**Risk**
Portable disk magnetic media	Diskette, Bernouilli	Variants on standards are common, but usually do not prevent retrieval	Prone to catastrophic damage
Portable disk magneto-optical media	Optical disk	Specialised products will become obsolete in the foreseeable future	Lack of data reading device
Portable CD optical media	WORM, Erasable CD	Variants in structures and formats	Loss of access to data
Portable Tape volumes	DAT	Variants in structures and formats	Loss of meaning
Network, server-based and Mainframe based	Disk drives, Tape reels, cartridges, MSS device	Volume and special operational environments	Loss of portability

Figure 5
Scorecard: Types of Media

Archive Requirements			
Feature	Capture	Storage	Best of available technologies (1997)
Longevity			
Viability	Must be able to be checked quickly for readability	Must use proven processes to refresh and restore	DAT
Obsolescence	Must be used as standard archiving medium	Must not use leading edge technology, must have proven durability	
Portability			
Price Performance	Reusable medium at little or no cost	Ability to easily back up copies for off-site storage	DAT
Ubiquity	Can be used as standard publishing medium	Can support access by many users simultaneously	CD
Susceptibility			
To physical damage	Not affected by stray magnetism	Can be held in racking	CD
To accidental damage	Can be sealed and self-contained	Can minimise data loss over time	

Figure 6
Scorecard: Recommended Archive Media

The recommended technology can only be that which is most suitable at the time. In the same as the Scorecard evaluates preservation candidate's technology profile, so will the archive technology be reassessed on a periodic basis.

Section 6. Dimension 4: Type of Platform / Operating System

In the last 50 years, the computer marketplace has seen every combination of personal, workgroup, divisional and corporate computing promoted as the answer to business problems. Each new machine range has adopted to make a step change in functionality in order to outdo the competition and protect its customer base. About 25 years ago, IBM halted its FS (Future Systems) programme because it had established through market surveys that it would lose half its customer base if it introduced a radical new technology that required everyone to change their programs and files. This same fear of losing market share still dominates the Operating System platform. "Transparency to the user" is declared for every major change in order to allay fears of another costly transition. Convergence is therefor in progress over a wide range of hardware platforms via the operating system and its "open" file structures and encoding techniques.

This is of great advantage to an archive. The hardware platform is not material as long as the archive media has an operating system-independent file recording and encoding structure. This is not the same as having a file which is ASCII, and can run on UNIX and Wintel platforms. Incompatibility can be hidden by product badges, for example, Windows NT 4 supports two file structures: the DOS structure which has a weakness for fragmentation, and NTFS which is not compatible with other Windows and DOS formats, because it structures the data on the disk to avoid fragmentation and consequent waste of disk and processor resources. In the same way compatibility of recording material is taken for granted today for CD, audio cassette tapes and videotapes, but it was not always so. Recent product developments are soon to disrupt the status quo again, both in the home and in the office.

The question of emulation has followed each major step forward in computer technology over the last 30 years. Hardware emulation has usually been preferred in order to provide speed and compatibility. Joint hardware and software emulation (sometimes on punched cards with the IBM/360 Model 25) has been used. At the present day, software emulation dominates, being programmable even at the chip level. In an archive, it may be necessary to handle some emulations, but this can only be tenable in the short term, while both the emulated and the host emulator are current in technology terms. Obsolescence for the host environment will bring double jeopardy for the emulated environment. Archiving of an emulation and its dependants should be considered only for the near term, and in the advent of destructive forces.

The study recommends that four platforms are suitable for consideration: Windows-base (primarily Windows '95), Windows NT, SCO-UNIX and OS/390. Any emulation should work within these environments.

Section 7. Evaluating the Preservation Requirements

The Preservation Complexity Scorecard

The final matrix scores (Figure 7) should reflect the level of complexity expected from the candidate digital material. The complexity factor represents, among other factors, the amount of human intervention that is likely to be needed, and this may only be apparent when all the factors are seen together.

Planning the Preservation Approach (Figure 7)

Matrix	**Matrix Result**	**Cross Matrix Complexity Issues**
1. Material	Score 1 - 5	Any score higher than 2
2. Formats	Score 1 - 5	Any score higher than 2
3. Media	DAT or CD	Any other media
4. Platform	One of 4 O/S	Any other Operating System

Score	**The Preservation Approach**
Matrices 1 and 2	
All scores lower than 3	Standard procedures in management of the archive will be sufficient to ensure long term preservation of the item
Any score over 2	Intervention required at the Capture stage, in order to (a) edit the item's format (b) remove parts of the item that do not need preservation (c) translate the item into an acceptable format, checking for loss of data (d) analysis of preservation requirements and the establishing of a special environment etc.
Other Matrices	
Not DAT / CD	Data transfer is required, extra cost involved
Not standard O/S	Data transfer and possibly data translation necessary. Checks necessary to ensure no data loss

Other Characteristics	Data transfer and possibly data translation necessary. Checks necessary to ensure no data loss
Age	If greater than 5 years, conduct trial capture to check for problems
Volume	If greater than 2 Gigabytes, conduct trial data capture exercise to validate estimates
Timing	Will the data be submitted in batches or all at once?
Delivery	Will delivery be electronic or by physical media?

Steps in the Preservation Process

Summary of Sections 8 and 9

Figure 8 is a representation of the functions within the management of an archive facility. The major issue arising from running a long term preservation service is customer confidence. When stakeholders entrust unique and possibly priceless material to the archive is a greater commitment than that made by most businesses.

The central six functions, stretching from Capture through to Access / Retrieval, are the core of the activity. The overall planning, reporting and administration are essential functions required to run the archive facility. The archive will need to be seen to be run as a business, particularly from the point of view of many of the stakeholders.

The archive is the goal for all the preserved material and the associated contextual data. How it is managed and run operationally will have an effect on the whole preservation process: efficient procedures during archive will not only reduce costs during Capture, but also further back up the business chain of activities, possibly improving the efficiency of the creators themselves. By increasing the confidence in the security of the archive and the integrity of the items, stakeholders may consider changing the way they work.

By managing the central process as a value chain, from Capture to Retrieval, improvements can be made in the way staff work together, controlling costs, improving service levels, and raising quality levels. The concept of project management will be very beneficial in managing all the resources more efficiently and controlling the changing technology environment so that the monthly operating schedule can be delivered on time, to budget.

Section 8 describes Figure 8 in more detail.

Section 9 relates the two functions of Capture and Preservation Engineering to the Scorecard and comments on the criticality of Storage Management

Section 8. Management of the Archive

Figure 8 is taken from a functional analysis approach, linked to Strategic Business planning. The approach is used to simplify the complexities of an organisation's working so that commonality of problems and differences in perceptions can be resolved without arguing about the meaning of a particular English noun. The model has four levels, which are described below in outline.

Strategy, Policy and Planning

The archive needs a forward view of where change will strike it next. A six-monthly review is sufficient for planning purposes. Technology trends take 2 - 3 years to unfold, and this would fit well with the Technology Watch which triggers off revision of the Scorecard. Apart from assessing the potential obsolescence of the new candidate digital material, the Technology Watch also permits more detailed transition planning to be made for the archive's configuration. Reports will also indicate where bottlenecks are occurring within the archive's current configuration and procedures.

Reporting

The finger on the pulse enables the management team of the archive to plan ahead in the short term. With information feeds from Accounts (Costs and Revenue), Help Desk (problem areas), Operations (virus detection, security), HR (resource utilisation), the team, will be able to direct and supervise the archive process month by month.

Administration Management

The archive will have all the requirements of a small business to manage its assets (its staff, its customers, its machine configurations), draw on local expertise (HR, legal and accountancy) and report back to branch management (and ultimately senior management) on the day-to-day practicalities and trends.

Project Management

The control of daily and weekly schedules is planned here, with an eye on quality and service levels. Because of the time perspective of the archive, standards are necessary in every facet of the operation, in order for there to be a consistent standard over a 50 year period.

Capture through to Access / Retrieval

It may be 30 years in happening but the data that is captured today will need to be managed so that it can be retrieved one day. Environment Engineering is the function that maintains the access paths and keeps technological obsolescence at bay. Taking its guidance directly for the Planning group, and advised by the Technology Watch, Environment Engineering makes the monthly tactical decisions.

Section 9. Capture, Pre-Preservation and Storage Management

The Scorecard for the candidate digital material is the major guideline for this activity. With sufficient experience and guidelines it will eventually be possible to plan the reception of the candidate material in advance, and convert the process almost into a production line.

The Capture team will form close relationships with the stakeholders who are submitting the material for preservation. In due course they may make a self-assessment of their material, and submit their judgement to the reception staff for advice and guidance. The Capture step is necessary to avoid substantial waste of time of major resources when the preservation is finally committed.

Preservation Engineering is the preliminary work that is necessary to differentiate which material is evidential and which material can be thrown away. Some ephemeral material (manuals or instruction books) may be duplicate or not required, in which case a photograph would satisfy the record. In the whole preservation cycle, it is at this point that loss may occur. Depending on the procedures invoked by the Scorecard, the material may go through a media conversion (copying), a format conversion (elimination of idiosyncrasies), material conversion (film into digital images) and processing conversion (alterations to the structure of the object and the way it will stored from now on). Just as the preparation is important, so is the testing that the material is now preserved, and cannot be interfered with. Quality assurance of the result is essential for good faith to be maintained with the stakeholders.

Feedback to the Scorecard is a useful function of the first two processes in the archive value chain. It will inform the guardian of the Scorecard of any changes that would help to make the pre-assessment more accurate.

All the careful practices of the pre-preservation team can be set to nought by the destruction of the archive copy. Every time the record is accessed, an opportunity exists for loss to occur. Storage management is generally about housekeeping, looking after the disk store, to ensure that nothing unplanned is happening.

Storage Management is very cost sensitive, therefore as little activity should be taking place as possible. Costs are incurred with every transfer, and with every intervention by an operator with a tape or disk.

The key movement of data will be associated with the need to refresh data, particularly on tape volumes. CD will not need refreshing, but systematic checks are a standard precaution. Storage Management is the key component in every management process, at some stage the data must be stored, retrieved, updated.

Section 10. Preparation for Change

Technological Change

The Scorecard is the means of monitoring the current acceptable complexity level of technology, and tracking those items that were preserved in the past with a different threshold. When this facility is allied to the principle of a Technology Watch, there exists a means to be prepared in all three steps in the Framework for technological change.

A Technology Watch on its own will do no more than act as a Cassandra, giving non-explicit warnings with no definite timing. When a technology prediction is linked to a database of existing archived items, the scale of the forecast and the potential impact, can give a manager some idea as to how to react and when to time the reaction.

In Figure 8, the stepping stone approach is an adjunct to the Technology Watch. When change is unavoidable, the most up-to-date technology may not be the most attractive. Instead, a less leading edge implementation may allow one to miss out a conversion, because it has less risk of failure and may be more adaptable.

Contextual Change

The four level context model proposed in the first Working Paper is a summary of what many other conference speakers and articles have discussed. Behind the principles are an attempt to match the progressive selection of information from day-to-day life as it is processed, so that only the really important distilled information is left. With the preservation of digital material we are preserving far more than any other society before has attempted to store.

David Bearman's model (Figure 9) with an axis for each of the four characteristics, converted here to a table has similarities to the same four levels. His model shows the information progressing becoming a valuable element in an archive as a result of being processed (by an instrument through to a domain), brigaded as part of the collective memory, and given a purpose becoming knowledge within the wider world. The products of his continuum are candidates for preservation. The process by which they have become candidates is valuable information in itself which will be used to create the contextual levels.

The Framework, through the Scorecard, also sees that technology will change the speed of the lifecycle and the various stages in Bearman's model, by making the "Act" faster, allowing more "Traces" to be captured and retained, and diversifying the number of "Instruments" that will process the data. The Framework would then propose that the demand for more preservation will grow, as the archive is flooded with more records with institutional meaning. Preservation will be needed to stay abreast of the accelerated use of information.

Section 11. The Context for JISC-funded Studies

In Figure 10, the remaining six JISC-funded studies are placed against the four-question framework structure, in order to give them some value from the current study. Where a study covers more than one quadrant, care should be taken that the study is not prejudiced by the influence of one or the other, but balances both topics.

Study 3 - An investigation of the attitudes of originators and rights' owners to the responsibilities of digital preservation. Working Paper 1 of the Study Report in Section 1.3 highlights that there are 10 stakeholders with conflicting interests. Figure 1.2 in the Working Paper shows them in interaction with the first level of the four contextual levels - Evidence should be the Object Description. In practice they react with all four levels. This dimension should be investigated as well.

Study 4 - A study of costing models for long term preservation of digital materials. The KISS principle in Step 2 of the Framework seeks to reduce the cost of capture and storage of items in the archive. Along with the concept that there is no such thing as benign neglect with digital material, it is possible that the cost models may be very different from existing library models, and that the substantial "tail" of on-going support may make some forms of preservation financially impossible to sustain.

Study 5 - A study of the three main methods of digital preservation:

a) Technology Preservation is generally seen as not being feasible, as a computer processing environment is almost impossible to preserve indefinitely as a working museum artefact. Experience in the UK and audit work has shown that it is prohibitively expensive, even when the systems it was supporting were very valuable.

b) Technology Emulation in this study Report is termed preservation of the operating environment, and is considered to be acceptable only in the short term, while the host environment is itself technologically current.

c) Information Migration is a procedural escape of the preserved information from technical obsolescence of the technical platform. The principles are well documented and practised as "Copy Management"

Study 6 - An investigation into the digital preservation needs of universities and research funders: Opportunities exist in this environment for economies of scale in the preservation costs. Equally the specialised variants of preservation demand (either high value material or bulk preservation) and capture (predominantly electronic) could create a cost-effective though non-standard model for long term preservation.

Study 7 - An investigation of progress already made towards permissive guidelines for digital preservation: short term short cuts may be false economies in the medium term. A risk analysis should be conducted on these guidelines in order to see if dangers exist. The KISS principle works from the other side of the problem, by lowering the complexity and the risk together.

Study 8 - Report on sampling methods and techniques for collecting materials, on the nature and extent of institutional electronic archives, and the relevance of current archival practice to digital preservation: the Scorecard could help monitor and track progress. Experience suggests that archival services approach the problem from the KISS principle perspective.

Study 9 - an investigation of post hoc rescue, or data archaeology, of high value digital material which cannot be accessed because the required IT environment is no longer available: In Working Paper 1, Figure 1.4, the post hoc rescue is described as being a two step, probably expensive process. The Technology Watch should pre-empt the need in the medium term. The Scorecard will maintain an inventory of the embedded, potentially obsolete, technology, which should cater for 99% of the problem.

Another study recommended for consideration by this Study is:

Technology Watch: both the Scorecard and the archive environment need to be kept uptodate, responding appropriately to shifts in technology. Whereas the archive is managed by IT professionals, who are guided by their technology suppliers, the Scorecard must search for step-changes and shifts in technology use across a much broader user community. The Scorecard must periodically be revised, cognisant of the creators of digital material and their fads and investment in technology. Best practice in the use of technology is relevant only as a benchmark against which to judge the amount of effort required when material is presented to the archive. The Scorecard and Technology Watch should therefore be managed and supervised by an independant body, that is interested in long term preservation issues. The ideal candidate is the National Preservation Office. Publicity for the project and request for information is probably best achieved by using a series of Web pages on the NPO's Web site. Apart from publicising the Scorecard standards, the Web pages could hold a self-assessment questionnaire, which browsing Web users would be encouraged to complete. In order to balance this self-selecting survey, we would recommend that a formal survey of 100 organisations world-wide should be conducted, over the InterNet, asking them periodically what their most common formats were and how they were using them. The amount of resource required for this exercise would be minimal, and it could be sub-contracted out. The results of the straw poll of browsers and the survey could be posted on the Web, itself initiating feedback. In this way a balanced view of current and past usage can be built up, and a more accurate scoring on the Scorecard can be maintained. In addition, a group of "Wise Practitioners", associated with the NPO, could be consulted on forecasting technology trends, in order to allow the NPO to plan for technology step-changes.

In all the JISC and related projects, we would recommend that time and resource are put aside to allow full collaboration of people from different disciplines. The long term preservation of digital material is a subject which lacks many of the attributes of the well-established skills and knowledge of parchment, papyrus, paper and film-based preservation techniques. Therefore an open-minded collaborative approach will be essential if the digital library and the digital archive are to be as successful and as valuable as our current Collections and their priceless holdings. Neither technologists nor archivists alone can solve the problem, there needs to be a concerted effort over the next three years in order to prepare ourselves and our parent organisations for the deluge of archival material that we know is coming.

WORKING PAPERS

Contents

Diagrams and Tables

Working Paper 1: Principles of the Proposed Framework

Summary of Working Paper Sections 1.1 to 1.4

The framework uses the following four principles to manage the issues arising from the long term preservation of digital material. The principles help to express the urgency and pressure to preserve, answering the four questions posed in Section 2 of the British Library Research and Innovation Report **50**.

Capturing Four Levels of Context (Section 1.1)

Digital material requires a greater degree of positive effort to preserve its meaning and context than most non-digital artefacts. Four levels of context are considered to be necessary, represented by different types of metadata. The levels are the Object (a description of the object itself), Object History (information associated with the storage and control of the material), Provenance (a record of the ownership, events in the material's history and its intrinsic significance) and Society (a definition of the material's contribution to cultural memory).

Managing Preservation (Section 1.2)

Understanding the practical issues of preservation allows a balanced programme of preservation work to be planned. Scarce resources can be focused on ensuring the permanence of the digital material and while maintaining a cost-effective environment for the long term.

Co-ordinating the Activities of 10 Stakeholders (Section 1.3)

The framework identifies ten types of stakeholder, who affect the way the digital material is preserved and managed in the archive. Traditionally the creator, owner and user are seen as the main stakeholders. With the long term digital archive, the number of stakeholders increases significantly. The interplay of the stakeholders will determine when and where the archive is established, how the archive and its contents are used in the future, and how successful in the long term an archive for digital material can be.

Managing Technologies in Continuous Transition (Section 1.4)

The management objective is that the archive environment must be kept current in technology terms in order to permit unconstrained, secure access to all items in the store. Within that environment, the underlying technology is being altered by step changes in the configurations, and the technology will never remain the same from one to the next. The preserved material will reside within an environment which is in a continuous state of transition. Balancing enforced change with continued assured access is necessary in order to provide a stable operating environment.

Section 1.1 Preserving a Digital Object and its Provenance

Over the long term, a preserved digital item can lose its meaning unless its context is also stored in association with it. Non-digital material, such as printed books, papyri and paintings usually carry within themselves sufficient physical and contextual information which allows them to be interpreted without a great deal of assistance. The context can sometimes be provided by the script, the language, the media, the style, and the signature. The item itself, because of its structure, may often act as part of the item's historical record. Alterations, additions and editor's annotations may be present as part of the item. Even a palimpsest or an over-painted cartoon can provide extra evidence for the item's provenance.

The type of contextual information for digital materials is similar, but usually the digital record is two-dimensional, requiring the history to be explicitly stated as the evidence for it cannot be deduced. Digital material is different also because the means of accessing, displaying and interpreting the physical record may no longer be operative because of technical obsolescence. Whereas the human hand recorded the manuscript, and the eye can read it 1,000 years later, the digital artefact may be unreadable within 10 years without the right equipment. Whereas the Etruscan language, Linear A and B scripts and Mayan stele may still puzzle experts after hundreds of years, within five years even the simplest digitally-encoded, compressed image may never again be accessible.

The framework recommends that four levels of contextual information is held for all digital material. This is represented in Figure 1.1, which describes the levels in terms of their focus, goals and possible metadata.

Contextual Level	Focus of Activity	Goal	Scope of Metadata
Evidence	Object Description	Permanence	Object / item attributes: the distinguishing characteristics about the item, index number, name, creation date, size, format, author, etc. e.g. Document Summary Information
Editorial	Object Histories	Durability	Manipulation: the Editors' record, refresh record, "migration" or transfer across media and between sites e.g. Operator's log
Provenance	Collection management	Significance	Holding record: history of ownership, roles and responsibilities of involved parties, position within a collection. e.g. ISAD(G) describing Fonds
Society	Interpretation	Cultural Memory	Intellectual: links to other artefacts, contribution to the development of society, as it is today in the future, and as it is today in the present

Figure 1.1:
Four Context Levels: Focus of Preservation Activity, Goals and Metadata

What is required is a means to bind this information together, in such a way that it can be treated as a unique object in its own right. Current technology allows the creation of a "bound" document, for example, using Windows '95 and Office '97 technology. The scale of the binding is under the control of the author, starting from association of external objects with a document to importing the objects and embedding them within the document. The objects can be any Microsoft Office-based product output. Macros can be used to enhance and extend this function further.

As long as the capsule of information that is created is not affected by in-built technological obsolescence, the approach should viable over the long term. Similar principles apply to Lotus Notes-based applications. By linking the document to an environment, owned and managed by an international de facto standards maker, one ensures as best as one can, long term accessibility of the capsule. Even if Microsoft and IBM were to be broken up, the new owners of the patents and copyrights would be sufficiently powerful to maintain upward and downward compatibility path for the capsule and its hidden components. Currently most major office software providers provide a very wide range of "launch", conversion or import/export facilities, in order to keep their clients reassured and loyal, and to attract new users, who are in the process of switching products.

Creating a capsule helps maintain the association of the context with the preserved item. A capsule does not in itself preserve the integrity of the contextual records unless another level of security is used. The "evidence" could be secured by the use of PGP or public-private key encryption. The capsule would then be locked, and could not be tampered with. Access would continue as usual with the public key.

Recommendations

A four level contextual approach, with data dictionary entry definitions, should be built in order to provide an information structure that will permit the successful retrieval and interpretation of an object in 50 years time.

A study should be established to explore the principle of encapsulating documents using the four levels of context, stored in a format, possibly encrypted, that can be transferred across technologies and over time.

Section 1.2 Managing Preservation

This study focuses on preservation, ensuring the long term safety (or permanence) of digital material. It is not possible to ensure the permanence of the media and the working environments, which are associated with the digital material. They are both an integral part of the digital items, but they may or may not be recognised as such, when the items are first received. It is part of the preservation process to manage both of them. We may be able to make the environment endure, but we may not be able to keep it unchanged over many years. The medium may become obsolete with a few years. The media and the environment need the attention of the preservationist.

There is an obvious difference between working to ensure no loss occurs, and working with the knowledge that loss will undoubtedly occur whatever precautions are taken. The difference emphasises the different ways in which we manage the items in the archive and the archive environment itself. The items we endeavour to make permanent, the archive environment we can only strive to ensure that it will endure.

At accession of a digital item, how much is stripped away, and treated so that it will endure in some shape or form, a surrogate, a photograph? How do we identify that which may be discarded in its current form, over the period of the life of the item in the archive? How should we handle amendments, additions, corrections to the base item, should these be amalgamated and preserved as well?

Because we are using four contextual levels it is straightforward to split off the base level - the object itself - from the associated descriptive material. It is unlikely that Levels 2, 3 and 4 will need to contain evidential material, although they may refer to it. It is a more complex decision which will allow the stripping away of the Presentation layer from the evidential material, leaving only the raw material to be archived, retrieved and redisplayed in mid-21st Century 3-D graphics. Additionally there may be some value in preserving the packaging of the digital material, similar to the dust-jackets of 1920's Legal Deposit material or an example of a CD, equivalent to an amphora exhibited as part of a museum's exhibition. A locked capsule and the original medium carrying the digital material need not be treated as unique items, as long as the technology which they rely on is extant.

With technology we are in a much better position to understand the interplay of the resources involved. Preservation of the technology environment is within our capability. It is distinct from the complexities of preservation of digital material which we cannot see and cannot inspect except by the intervention of technology. By limiting preservation activity to a specific remit, it is easier to track technology trends for the technical obsolescence of the archive in the medium term. The threat of destruction is lessened.

Recommendations

Institute a Technology Watch for impending obsolescence of archive environments.

Apply the Technology Watch results on the Archive environment inventory on a 6-monthly basis. Act on all anomalies.

Should every preserved item therefore contribute to an inventory of the environments preserved within the archives? Such an inventory would record special processing requirements of specific items in the archive, in anticipation of technical obsolescence in the future.

Section 1.3 Stakeholders and Preservation Issues

There are ten stakeholders involved in the Preservation process (Figure 1.2). The stakeholders may be different people, or a combination of individuals and organisations. The Creators, Owners and Providers are the major stakeholders, recognised as being essential for co-operating with the Libraries and contributing to the holdings in the archive. Often at least one of these parties will have a financial interest through the copyright on the material. It is this interest that some archivists hope may encourage funding to be made available (from the Fund-Holders) in order to manage and maintain a digital archive in the long term. The Regulators have set the scene with legislation to preserve ownership for a limited period of time, to ensure a national collection of material is established and to preserve items that are in the public interest.

The situation changes with the plan for long term preservation. Over a long period, copyright will lapse, will this make it less attractive to the copyright owner to contribute? The Regulators may therefore have a role in extending the legislation to make anyone who uses an archived item to contribute to its preservation. The fee would be collected by the Provider, a re-publishing fee. The original Providers of the designated archive material are very likely not to be the Providers of the copy from the Archive.

Recently preservationists have identified a new Stakeholder: the Interferers. These individuals and organisations are frequently the antithesis of the current Regulators. New regulations are sometimes subsequently formed because of these pressure groups. At other times they are seen as simply a nuisance, obstructing the course of good preservation practice, taking a narrow perspective on minor issues. Their impact is usually to delay new measures and to initiate a review of current procedures. Interferers can be put to good effect by judicious lobbying.

Technology is an Interferer. The development of new media is driven by the potential revenue from accessing data, communicating information and developing commercially profitable knowledge bases. Storage is seen as a temporary issue, only a small proportion of the business event information is finally stored. The emphasis is on processing the data, in new and inventive ways, displaying the information as fast as possible anywhere in the world, linking various knowledge bases dynamically, and capturing more and more diverse data items to feed into the system. Permanence is not in the developer's vocabulary. Durability of a database is linked to commercial justification, not to the maintenance of a national archive.

The stakeholders will also have a contribution at the other levels of context. Budget cuts and political instability are Interferers at the Provenance and Society context levels. A budget cut can seriously damage the value of a collection, by restricting intake and causing holdings to be disposed of. A war can destroy centuries of preservation, the intellectual heritage of a culture.

Stakeholder	Activity	Impact on the Long Term Preservation of Digital Material
Initiators	Collection developers Risk Assessment, Technology Watch	Research libraries collect material that is current, published on current technology. Establish the nature and scale of the threat of irretrievable loss for digital material items
Regulators	Legal Deposit, Public Record Office Copyright	Assess current legislation to cover contribution to cost of conserving an archive
Creators	Record	No control over format of deposited items leads to unmanageable diversity
Owners	Maintain Copyright	Preservation of material will lead to demand for copyright in perpetuity
Fund-Holders	Financing preservation activity	Manage the funds available for preservation activity according to agreed priorities and service levels
Providers (1) - at embargo date	Publish	Initial diversity of formats at publication complicated by new editions in new formats and on new media. Archive copy should be deposited in an independent format
Readers / Access	Obtain copy of item (for a fee)	Readers will demand material in current acceptable format for display and inclusion in new digital material
Archivists	Refresh medium	Conserve the archive, whilst preserving the items, and maintain the integrity of the deposited items, against hackers and viruses
Providers (2) - long term access	Re-format onto new medium	Provide new editions, which link into the new intellectual context through re-indexing and re-packaging
Interferers	Make material inaccessible through technological turbulence or block publication	Technological progression is driven by use (processing and display) not by long term storage Pressure groups may cause some material not be published, or not stored or to be deleted from the holding

Figure 1.2:
10 Stakeholders: Activities and Impact

Recommendation

A more detailed study should be made of the inter-relationships of the ten stakeholders, and how they can be made to support the long term preservation of digital material. This will be linked to the economics of archive management (the cost model), changes in legislation (Legal Deposit, etc.), the risks of relying on links between National Libraries to maintain collections (threats of wholesale destruction of collections), and loss through viruses (technological turbulence).

Section 1.4 The Technological Long Term

Unlike the traditional archive, digital material cannot be the subject of benign neglect. With use or lack of use, digital material steadily loses its value, unless the item is actively preserved, and its environment is actively conserved. By definition, there is no long term technology.

Technology makes preservation of digital material difficult for five reasons:

First, if a digital item is captured today, its components will represent a legacy of technology, possibly from the last five years. An item can be assessed as to the age of its components, Figure 1.3 is an optimistic assessment of an object's technology content.

Presuming that the item was captured (written, edited, scanned, composed) on the latest equipment, it is likely that less than 5% of the total is represented by 1997 technology, for example, bug fixes. The latest Microsoft suite of office software (Office '97) will contribute 50% of the technology legacy, but it will be of 1996 vintage, which was when it was tested, possibly on advance shipments of the new hardware. The rest is mainly 1995 (35%), being standard core routines from Windows '95, unchanged by Office '97. Finally, elements of the base MS-DOS operating system (DOS version 7, and DOS emulation code) will remain embedded in the architecture of the PC system, this may still account for perhaps 10%. In contrast, Windows NT and OS/2 were written without any progenitors, and have a completely different composition.

If the same document, image or spreadsheet were captured in March 1998, the proportions would have changed, particularly if the hardware and software configuration had been kept up to date. In the main, however, many PC users are using software which is based on a platform which is pre-1995. Because it is suitable for their purposes, is reliable, at least with known glitches, they have made no attempt to change the basic configuration, adding components when required, year by year.

Second, even the concept of "migration" is not adequate to describe the changes that are occurring in every aspect of technology, hour by hour. We are using hardware and software components that are in a continuous state of transition in our office systems. Compaq built up their reputation by guaranteeing that the internal construction of their PC does not change, whether you order 10 or 100 from stock. Many other suppliers deliver a varying internal configurations for the same model, which causes many problems during upgrades, maintenance and trouble shooting. Software fixes are embodied within the next release of office packages as they are shipped, and the new configurations are rarely announced. We are using these "chameleon" PC systems to record critical aspects of our culture. The trend is to ever more complex technical implementations, easier to use for the user, but hiding increasingly complex interactions on the inside. Backwards and forwards compatibility is limited to formats, which may allow data to be rescued from obsolete systems.

Third, everything is “old”. It is not possible with any certainty to say that an object is “up to date”. This is experienced daily by InterNet users, because the user community, the connections, the sites, the links and the data can all change while one is searching. The concept of keeping a “master” version or copy requires corporate standards and controls. Keeping a management trail of the changes to an object requires a logging and tracking system, roll-back recovery facilities and mirroring of transactions. Synchronising the update activities of different systems, so that a consistent, up-to-date picture can be maintained is expensive, and tends to be limited to a few applications. Therefore it is more economic to assume that everything is out-of-date, initiate a search for the updates, only when they are needed, and manage the updates in the correct sequence as they become available from the search.

Fourth, digital material is currently preserved most easily by making many copies of an item. It is said that every letter ever written on a networked computer is stored somewhere, it has been copied as part of standard backup routines. The difficulty is no-one would know where to find it.

Fifth, the technology can be seen as a series of stepping stones to the future. In Figure 1.4, this stepping stone approach to conserve valuable items is described in diagrammatic form. It shows also that the technique, known as post hoc rescue, may need two steps in which to recapture data from an obsolete technology platform.

In summary, technology uses an implicitly different timeframe to the accepted principles of preservation, the “technological long term” has a very close or near horizon. The technology carrying the candidate digital material can be obsolete before most archivists would have started to consider conserving the items.

Recommendation

A technology management trail (within the Scorecard - see Step 2 of the Framework) should be established before the more complex digital material is stored. This is to ensure that, for an item of digital material, the full extent of the internal inter-relationships are understood, and the implications for long term preservation in a variety of successive environments are documented.

Working Paper 2: Issues Concerning Access in the Long Term

Summary

The perspective of 20 years from now is not one that many people take. Managing the archive requires that staff take the long view and work out how their actions may be affecting the health and availability of the data for future stakeholders.

The Procedures for Preservation dictated that material had to be treated consistently when being received into the archive. The Scorecard was used to establish a reference point for all capture and pre-preservation processing, and for it to provide a means of evaluating the scale of change and the impact of obsolete formats within archive items.

The continuously changing world of technology requires particular attention to the management of the storage technology. Without it, the archive would be lost, either through obsolescence or through negligence.

Taking something out of archive has its risks, locking it back into an operational environment may be more complex than the Scorecard originally made out and viewing it as one did when it was deposited in the archive may be practically impossible. These issues are briefly discussed and the issues flagged for discussion in the paragraphs that follow.

Retrieval of Preserved Items

When opening the capsule, the same precautions should be taken as if there were unknown material held within it. The Scorecard record gave some indication of the provenance and value of the material, but the techniques used to evaluate and check for glitches may not have been as sophisticated as they are now (2017 C.E.). By taking out an item form the archive and loading the document, one is taking a risk that a time-encoded virus is let loose at the same time. Therefore the first step must be taken within a security-firewalled environment, where tests can be made on the material.

The testing regime is not only a precaution against contamination but also a means of testing the locks the material may have been given to link in through the InterNet to re-establish its knowledge base. The process may be trail and error, as many of the links may have gone, or may have been upgraded, so that they are unrecognisable. The testing environment will be prepared for this and gradually the recovered material will be ready for its functionality test in the New World.

It is impossible to imagine what the test environment's reaction would be to an Office '97 application, and how pedestrian it may seem. Either way the gradual de-layering of the four context levels, and the progressive testing, will gradually enable the material to be put into its new context.

Reprocessing of Items

The preserved items will be moved into the new working operational environment, this may involve a certain amount of conversion. Agents would reconstruct indexes, tables of contents, and establish a new set of preferences. The result must be tested to see if it conforms to what the item looked like 20 years previously. In order for this test to be run properly, a thumbnail or some form of test result must be necessary in order for that the user who asked for the item to be retrieved will knows that what they are getting is what they expected.

This implies that the second level of context - the Editorial level - should contain some test data, against which one can set the expected outcome. With "bound" documents, this Editorial level may have to contain several testing databases.

Redisplay of Items

Because the redisplay capability of machines five years ago is now so far separated in terms of functionality and price, we would not consider trying to imitate the displays today. There has to be a decision made as to how far we pursue the purity of colours and lines per inch, and conspire to produce a replica image.

In 20 years time, there will be 3-D user interfaces, with automatic format conversions and agents that establish new links with databases with the same interests as the document. Just as the original Lotus 1-2-3 spreadsheets cannot now be displayed, one has to wonder is there any purpose in keeping it in that format? It would work equally well in Excel. What is it that we are preserving?

Republishing in 2017 will use an entirely different approach than we have today.