Perfect one day – digital the next: challenges in preserving digital information

ALAN HOWELL

Abstract Preserving information is about maintaining its meaning – or essence - over time. It is achieved through maintaining information's usability, context and content. This article compares and contrasts some of the characteristics of information in paper-based format and information in digital format, particularly the characteristics that affect their preservation. The leading ideas on why digital information is inherently short-lived are discussed along with current-best thinking on how digital information may be preserved and some recent initiatives in Australian libraries in this area.

Our information-loaded, some would say *overloaded*, world is living in troubled times. This is not surprising when just over half the world's English-speaking population, and just under half of Australia, is reputed to be online.[1],[2] Then there are the estimated seven million pages that are added to the Internet each day.[3] Arguably, an exponential explosion of digital information is overwhelming not only business and personal life, but also the capabilities of libraries to fulfill their traditional functions.

The life expectancy of digital information is also problematic. It may be as short as Brewster Kahle's estimate that the average lifetime of a URL is 44 days.[4] Concern is growing in the library literature. Two examples set the tone. Kahle and Peter Lyman, who jointly run the Internet Archive, have written that 'digital information has been allowed to become a medium for the present, neither a record of the past nor a message to the future'.[5] Terry Kuny, a consultant to the UDT Core Programme at the National Library of Canada, thinks that 'we are living in the midst of digital Dark Ages'. Kuny echoes Kahle and Lyman when he writes 'that we are moving into an era where much of what we know today, much of what is coded and written electronically, will be lost forever'.[6] These are ominous predictions, with potentially serious consequences for libraries and their customers. Even if they are partly realised, failure to recognise and respond to these concerns may result in individuals losing all or part of their personal histories; business, industry and government losing valuable intellectual property and the ability to govern; and an unfillable hole appearing in the collective memory of our society.

What has caused these negative perceptions to arise and what are we doing to address these concerns? One answer is to look at the information resources in libraries and how they are changing in a digital world.

Traditional and contemporary information resources: what's the difference?

Paper makes perfect

Libraries contain mainly printed books, hand-written manuscripts, archival records, maps, photographs and works-of-art-on-paper such as watercolours. For the purposes of this article I am going to collectively call them *resources*. They are analogue (continuous) materials mostly made from paper, a wonderful material on which to print or write information that is intended to last and be easy to use.

Paper's ubiquity should not be surprising. We have nearly two thousand years experience in making paper by hand and, for the last two hundred years or so, making paper by machine. When kept in cool and dry conditions, high quality paper that is also moderately alkaline and made from purified cellulose pulp is likely to remain flexible, durable and last for several hundred years. Due to its demonstrated permanence and durability over time, the life expectancy of high quality paper has become the benchmark for information resources of enduring value.

Being digital

In contrast to the continuous analogue resources described above, the word *digital* means discrete. Digital resources are encoded in a pattern of signals that must be in only one of two logical states; either one or

off, either one or zero, or either high or low (depending on the industry in which you are working). This binary simplicity - and the common means of storage, transmission and visualisation that it produces - is both the overwhelming strength and the fundamental weakness of digital information resources. The first strength of being digital is that affords the potential of multiple simultaneous uses from a single original in ways that are not possible for resources in any other format. The second strength is that digital information resources have the potential of universal access. This advantage is tempered by the numerous incompatible encoding systems and the significant cost, language, location and speed restrictions to be overcome. The final strength is that every digital resource has the inherent potential to be preserved in perpetuity. I will expand on this unique characteristic in the next section.

Digital information resources include information resources that have been digitised from an analogue state and digital information resources that only exist in the digital domain, the so-called *born digital* resources. The latter include virtual reality simulations, computer-aided design (CAD), geographic information systems (GIS), economic models and real-time interactive World Wide Web sites. Floppy disks, compact disks (CDs) and digital audio tapes (DAT) are familiar carriers of, and storage media for, digital information. Along with virtual digital information resources such as the Internet they are used nearly every day in jobs and leisure activities.

The preservation perspective

From a preservation point of view paper-based resources deteriorate slowly. In controlled environments such as library storage this deterioration is almost imperceptible. In contrast, digital information resources deteriorate rapidly. Carriers of digital information can be made to last many years, however magnetic materials and signals are far less stable, possibly as short-lived as one year.[7]

Paper-based resources deteriorate internally and chemically, at least to start with. Physical deterioration only becomes apparent when total deterioration is well advanced. It then takes place rapidly. Another way of putting this is that paper-based resources have a very high redundancy. In contrast, physical electronic media are rotating technologies and subject to manufacturing defects, inherent faults, wear and tear, damage and disasters. Virtual electronic media, and all electronic signals, are subject to corruption, interference from stray magnetic sources and signal decay. This deterioration is not only highly localized but also difficult to *see*. Even the corruption of a small number of signals maybe sufficient to render a digital resource unusable. Digital information resources therefore have a very low data redundancy.

Although it may not seem at first glance to be an obvious characteristic, paper-based resources must be managed with their carriers. The print on the page of a book cannot be separated and handled independently, for example. In contrast, digital information resources must be managed separately from their carriers, which are rarely designed to be long lasting.

Paper-based resources are usually managed reactively and retrospectively. Almost all the paper-based resources that we are preserving in our libraries are years, if not decades, old. In contrast, digital information resources must be managed proactively and contemporaneously. Due to their rapid technological obsolescence, which is discussed later in this article, it appears likely that there is only a three to five year *window of opportunity* in which to make this decision.

The contrast between paper-based resources and digital information resources is also evident in the discussion about their *metadata* (the underlying definition or description of the data in a resource). Books, manuscripts, files, newspapers and photographs for example are easily recognised as what they are. Paper-based resources may therefore be said to carry with them their own metadata. In contrast, digital information resources require externally provided metadata, as they cannot be recognised by simple observation. This raises the interesting question of how to digitally preserve metadata, which cannot be *read* without a higher level of digital metadata without yet another higher level of digital metadata, etc.

Paper-based resources have the potential for high artefactual value. The preservation rational is to conserve the artefact. A whole conservation industry has been built up in libraries to do just this. In contrast, digital information resources have no artefactual value. The digital preservation rational is to

preserve the information and discard or re-use the artefact. Re-usable digital storage media such as floppy disks and rewritable compact disks are the best examples of this characteristic.

Another important discussion is about the cost of preserving paper-based information resources compared with the cost of preserving digital information resources. The received view is that paper-based resources cost less to preserve over time. This is mainly due to the low unit cost of air-conditioning and the very slow rate at which most paper-based resources deteriorate. In contrast, digital information resources may cost more to preserve over time due to the need to *refresh* and/or *migrate* them to new carriers and/or new systems every few years. However, until a critical mass of digital information has been preserved for a decade or so, this aspect of this discussion is very difficult to comment on.

The final characteristic is that paper-based resources are used up *a little* and lose a little of their information content each time they are used or copied. In contrast, digital information resources need to be used to exist. Each use of a digital object refreshes its signal and potentially subjects it to error recognition and error correction systems, if these are present. This unique characteristic has caused Steve Gilheany of Archive Builders, a US company specializing in manual and digital corporate archives and records management, to write that 'we now have the promise of preserving information forever'.[8]

The differences between paper-based information resources and digital information resources are significant yet easily overlooked. Preservation of the former category is well understood and widely practised with generally satisfactory outcomes. Preservation of the latter category is not yet fully understood and consequently not yet integrated into library preservation programs.

Problems with pixels

The life expectancy of digital information resources is problematic due to their machine dependency, unpredictable media life, rapid technological obsolescence, and reliance on us for their survival. That is there are significant management and sociological issues.

Machine dependency

Digital information resources offer considerably increased functionality. This is offset by their increased complexity. My work, and no doubt all our yours is now almost completely dependant on access to a personal computer; a concept that's barely two decades old. Using and preserving machine-readable formats by definition requires the information, the medium and the access technology to be in good condition. Without compatible and operational hardware, software, probably a telephone and certainly electricity, information in digital format is incomprehensible, useless and irrelevant.

Unpredictable media life

Although all information storage media will degrade eventually, the life expectancy of digital information resources depends to some extent on the medium (or carrier) on which they are stored, at least in the short-term. Although any digital preservation strategy has to consider the longevity of digital media, building a strategy on conserving digital artefacts such as floppy disks, computers and other hardware, is a largely pointless exercise. Nevertheless, our short-term reliance on digital carriers depends on six factors:

- the quality with which the medium was manufactured
- the number of times the medium is accessed over its lifetime
- the care with which the medium is handled
- storage temperature and humidity
- the cleanliness of the storage environment
- the quality of the recorder used to write to the medium.[9]

Early attention to the difficulties in preserving digital information focused on the longevity of physical electronic media. Surprisingly, this is the subject of much discussion and controversy. Not surprisingly, manufacturer's estimates and experts' estimates widely differ. A few years ago Mark Rothenberg, a senior research scientist from the RAND Corporation, wrote in *Scientific American* that the guaranteed lifetime

of magnetic tape was one year.[10] At the same time Kodak® were reporting a '100-year life-time at room temperature' for their optical media.[11] Both estimates are still well short of the life expectancy of high quality paper.

Magnetic tapes With moderate care, most magnetic tapes used for digital data storage will last for ten years. With special storage and handling, digital magnetic tape formats can reliably store information for thirty years or more. Storing tapes at lower temperatures and humidities can significantly increase life expectacy.[12] However, whatever the longevity of current digital storage media, of greater significance is that life expectancy has to be set in the context of recording and playback technology.

Optical disks For optical disks of all kinds, accelerated testing appears to be showing that their life expectancy is greater than the life expectancy of the technology on which they are dependent. Drawing on research in the technical literature, particularly by John Van Bogart, the instability of magnetic tapes and optical disks is elegantly summarised by Ross Harvey in *From Digital Artefact to Digital Object*. Harvey concludes 'that there are at present too many unknowns to commit digital data to currently available artefacts for anything other than short-term storage. The preferred option is to direct preservation efforts towards solutions which preserve the information content - the digital *object* - rather than the digital *artefact'*.[13]

Technological obsolescence

In today's fast-moving, market-driven, highly competitive, convergent and electronic world, hardware, software and operating systems for recording and storing digital information are essentially obsolete every eighteen months.[14] Described as *technological obsolescence*, this is the most pressing technical issue and is a far greater threat to information in digital form than the inherent physical fragility of all digital media. Kuny gets to the heart of this issue when he writes that 'the challenge in preserving electronic information is not primarily a technological one, it is a sociological one. The dynamism of the market for information technologies and products ensures the fundamental instability of hardware and software primarily because product obsolescence is often key to corporate survival in a competitive capitalist democracy'.[15]

Sociological and management issues: bits know no borders

Compounding the technical challenges of preserving digital information is the challenge of accepting that there is a challenge to be addressed. In a *throwaway* digital culture, where new versions of software and hardware regularly usurp yesterday's *latest version* – with vary degrees of backwards compatibility - the concept of digital preservation is an anathema to the general community. Other challenges include reaching agreement on the key terms used in the preservation of digital information resources, coping with rapid organisational change, a lack of standards, managing digital intellectual property, addressing the concerns of multiple stakeholders, deciding what to preserve and deciding who does the preserving. These considerable challenges are now described in more detail.

Reaching agreement on the key terms used in the preservation of digital information resources Harvey illustrates the issue of terminology when he discusses the term *archival, which he says '* appears to have many meanings'. To archivists and librarians *archival* 'means a life span of several hundred years'. In contrast, 'manufacturers of compact discs talk in terms of decades', and 'computing people may talk of up to two years'.[16]

Coping with rapid organisational change The management of digital information resources requires the organisational ability to accommodate rapidly changing digital technologies and the capacity to identify, acquire, preserve, and provide user access to these ever-changing resources. The challenge here is that as librarians, archivists and conservators we have built our professional rationale on standards-driven and painstaking incremental advances over many decades. We do not have the time to do this in the digital world.

Addressing the lack of standards In contrast to the preservation of paper-based information resources, the preservation of digital information is largely experimental and fraught with the risks associated with untested methods. Although electronic resources have been in existence for over sixty years, there is an absence of established policies, standards, protocols and proven methods. In their absence, *de facto* standards, such as HTML and PDF are being widely adopted for the dissemination of documents and TIFF for the storage of images. We need to work to create their equivalent *de jure* forms.
Managing digital intellectual property The challenge is to resolve the widespread uncertainty about the legal and organisational requirements for managing digital intellectual property. At present, text and other document-like resources, photographs, film, software, and multimedia resources each operate under very different regimes.
Addressing the concerns of stakeholders Stakeholders in the preservation of digital information resources include information creators, owners, managers of digital archives, representatives of the public interest and public policy, and actual and potential users of digital information. In general, providing access to and the preservation of paper-based resources has been left to professional archivists and librarians. This is largely because paper-based resources in institutions have lost their commercial value and archives and libraries are trusted repositories. The perception is that they are unlikely to seek commercial or political profit from the ownership of these resources. This is not the situation for digital resources, which are likely to be commercially valuable and may also be politically sensitive. The challenge is to understand and protect the concerns of stakeholders while ensuring preservation and access objectives are also maintained.
Deciding what to preserve. This is an issue to which our traditional response has been to keep everything that we can and wait for the significant information to rise to the surface as time, people and events dictate its importance. Projects in the digital domain taking this approach include the Swedish Kulturarw[3] project, the Finnish EVA project, and the Internet Archive; a project 'to collect and preserve the entire *web*, past present and future'.[17],[18],[19] Realistically, this question may well now have to be rephrased as *what can we afford to preserve?* as the cost of preserving digital information resources rises at a time when funding for libraries and archives declines.
Deciding who does the preserving. The optimistic aims of our institutions to accumulate comprehensive collections may have to be tempered by the reality that a great deal of the interesting digital information is not in the public domain. Kuny comments that information is becoming increasingly commodified and 'companies will be the place where the most valuable information is retained and preserved, and this will be done only insofar as there is a corporate recognition of the information as an asset. But companies have no binding commitment to making information available over a long term'.[20]

Preserving library futures?

When it comes to preserving information in digital format, the aim is to preserve meaning over time through maintaining usability, context and content: the intellectual substance contained in all information resources. This is a complex idea that operates at several different levels of abstraction. Preserving meaning, so that the ideas and facts contained in a resource stay identical to those contained in the original resource - irrespective of format - is more important than preserving either physical or virtual carriers or the exact number of *bits* and their precise order. Hilary Berthon and Colin Webb have also described this as preserving *essence*, which they define as 'that must be preserved, and procedures for authenticating its survival over time'.[21]

Although complex, time-consuming and costly, preserving meaning brings with it significant opportunities to re-engineer and leverage information resources for their enhanced service, product development and wealth-creation potential.

CPA/RLG Task Force

A major contribution to the thinking on this issue in the library sector was the publication in 1996 of

Preserving *digital information: Report of the task force on archiving of digital information*, by the U.S. Commission on Preservation and Access (CPA). The report arose from a decision at the end of 1994 by the CPA and the Research Libraries Group (RLG) to create a Task Force charged with investigating and recommending 'the means of ensuring continued access indefinitely into the future of records stored in digital electronic form'.[22]

The Task Force was charged specifically to frame the key problems (organisational, technological, legal, economic etc.) that need to be resolved for technology refreshing to be considered an acceptable approach to ensuring continuing access to electronic digital records indefinitely into the future. The Task Force was then asked to define the critical issues that inhibit resolution of each identified problem and, for each issue, recommend actions to remove the issue from the list. Finally, the Task Force was asked to consider alternatives to technology refreshing and make other generic recommendations as appropriate.[23] The Task Force's conclusions were:

- Creators and owners have primary responsibility for the preservation of digital information.
- A *deep infrastructure* capable of supporting a distributed system of digital archives is required.
- Storing, migrating and providing access to digital collections depends on a sufficient number of trusted organisations.
- Certification for digital archives is needed to create trust.
- Certified digital archives must have the right and duty to rescue digital information if its current custodian is negligent.[24]

The Report, and more importantly the inclusive process adopted by the Task Force to write it, triggered other preservation communities to contribute to the process, learn from it, and work with their own constituencies. In this respect Australia can be proud of its own contribution.

Action in Australian libraries

In Australia, and particularly through the exemplary work of the National Library of Australia (NLA) the library preservation community has taken a number of innovative and internationally well-received collaborative initiatives. They include the *Preserving Access to Digital Information* (*PADI*) project; the *Preserving and Accessing Networked Documentary Resources of Australia* (*PANDORA*) strategy, and developing seven *Principles for the Preservation of and Long-Term Access to Australian Digital Objects*. These initiatives are now described in more detail.

PADI (http://www.nla.gov.au/padi) is the NLA's subject gateway to resources about digital preservation. It has an associated discussion list, *padiforum-l*, for the exchange of news and ideas about digital preservation issues.[25] Currently, the PADI web site contains records pointing to approximately 850 resources in print (1%) and electronic format (99%) organised in an underlying metadata repository. The browse interface is refreshed dynamically from the repository and an interface enables users to search the repository using a range of criteria, including fielded searching.

PADI describes quality digital preservation resources regardless of format. Where there is no direct access to content, users may obtain further information about the availability of the resource through the repository record. PADI uses the Dublin Core metadata schema in all processes involving interoperability with other systems. The full PADI metadata schema is described on the PADI site.[26] PADI is being developed as an international subject gateway. To ensure a high level of quality and relevance to the subject matter, the resources linked from the gateway go through a selection process managed by the NLA with the help of an international advisory committee.

PANDORA (http://www.nla.gov.au/pandora) is a strategy to provide long-term access to significant Australian online publications. In the PANDORA strategy online publications are identified and catalogued onto National Bibliographic Database. Arrangements are also made with the owners of identified publications to capture a copy of their publication for safe keeping in the PANDORA electronic archive and make the information from the archive available to users in line with fair dealing provisions

while taking into account publisher's commercial interests. Backing up this strategy is the technology to preserve the *look-and-feel* of online electronic publications wherever appropriate, update the information in the archive on an ongoing and systematic basis, while maintaining date stamped previous issues or versions, and migrate information in the archive to new formats as current technology ages.

The NLA considers that the preservation of all significant Australian Internet publications is beyond the capability of any single agency and that a cooperative effort among the traditional deposit libraries is essential. The State Library of Victoria formally joined the PANDORA Project in 1998. They have produced their own set of selection guidelines and are selecting, cataloguing and archiving online Victorian publications. The State Library of South Australia and ScreenSound Australia have also recently joined the project and will soon begin selecting, cataloguing and archiving online publications in their areas of responsibility. The State Library of Tasmania is not a member of PANDORA, but is an important contributor to the wider Australian cooperative archiving effort. They have their own project in hand to archive Tasmanian online publications, *Our Digital Island*.[27] The State Library of New South Wales has also been experimenting with their own system for archiving selected online titles.

PANDORA is building towards a distributed national digital archive. As it is conceived so far, the national model would consist of a network of distributed archives, with the NLA and the state libraries working to an agreed set of principles and actions, and gathering the titles for which they accept responsibility into an archive maintained within their own institution. While each library may employ different internal procedures and technical platforms, the basic components of the national model are a set of collection agreements which will delineate the areas in which each participating library will take responsibility for archiving Australian online titles; a commitment from each participating agency to catalogue all titles selected for archiving onto the National Bibliographic Database, which will be the principle means of access to the national collection of Australian electronic publications; commitment from each participating agency to undertake the necessary steps to ensure long term preservation of the titles for which it has accepted responsibility; and agreement from all participants to negotiate arrangements with publishers that will ultimately ensure open, networked and gratis access to titles in the archive.

Statement of Principles for the Preservation of and Long-Term Access to Australian Digital Objects (http://www.nla.gov.au/preserve/digital/princ.html) In 1995, the Australian Cultural Ministers Council endorsed its *National Conservation and Preservation Policy for Movable Cultural Heritage*.[28] Meeting shortly thereafter, a cross-sectoral workshop on digital objects developed a Statement of [seven] *Principles for the Preservation of and Long-Term Access to Australian Digital Objects*. The seven Principles are:

- The cooperation of all interested parties is essential.
- The creators of digital objects have the initial and in some cases a continuing role in preserving access to them.
- The location, selection, identification/cataloguing and retention of digital objects will be best achieved through the coordinated distribution of responsibilities.
- Access to digital objects should be preserved only for as long as they are judged to have continuing value and significance.
- The rights of creators, owners, providers, users and subjects of digital objects must be balanced and protected.
- The adoption of best practices and standards is essential.
- Appropriate Commonwealth and State government regulatory, legislative and policy regimes are essential..

These seven operating principles broadly underpin current worldwide initiatives to preserve digital information resources including the PADI and PANDORA projects described above.

Getting technical

At the technical level, the preservation of digital information resources in digital format for archival, business-continuity or counter-disaster purposes involves careful handling and storage, *refreshing* (copying), *migration*, *emulation* and the preservation of meaning (or essence). These strategies are now described in more detail. There is considerably more information on the PADI web site. Printing digital information resources to paper or *writing* them to film and *digital archaeology* (techniques for recovering information on obsolete carriers or in obsolete system) are also briefly discussed. The techniques all involve the separation of digital information from their carriers, which are considered to be disposable.

Handling and storage Although only a short-term preservation strategy, the initial management of digital information resources depends on its careful handling and appropriate storage. Although this strategy is used to preserve all documentary heritage the following ten recommendations for preserving digital information may be helpful.

- If you are the creator of the digital information resource then make preservation of it easier, either for yourself or for someone else. If you are using one of the many widely used data formats such as .doc, .htm, .pdf, .ppt, .tif and .xls, then these formats allow creators to embed descriptive metadata. This is your opportunity to leave a record of creation, ownership and technical characteristics that has the potential to move forward with the information as it is preserved. In addition, if you have a choice of format to use, then choose a *de jure* or *de facto* standard format to work with. Digital information resources created in standard formats will have an increased chance of being preserved. This is because there will a critical mass of information to be worked on. In turn, this leads to a market for the preservation of standard format information and consequently lowers unit costs.
- Choose digital media with preservation in mind. All digital media are not created equal. Manufacturer's catalogues and web sites are a good source of information on their products. State and national institutions, particularly those that specialise in the collection of digital resources, are also good sources of information on the better carriers to use.
- Make a record of the technical characteristics of the information and the media. These may come in handy when you want to copy them.
- Keep representative examples of the original packaging and any information leaflets supplied with the media. The information is there for a reason, and it may be useful. Michael Lesk, Division Director, Information and Intelligent Systems at National Science Foundation, U.S., has mused that 'we will need a new profession, to be called perhaps *digital paleographer*, to decipher the formats of out-of-date packages'.[29]
- Follow manufacturer's recommendations about use or storage.
- When copying information move it to higher, not lower, quality media.
- Handle with care. Digital storage media are not as robust as they appear.
- Although it is beyond the scope of this article to give precise recommendations about environmental conditions for the storage of digital information resources, cool, dry, and dust-free conditions are likely to promote longevity.
- Test systematically to establish the condition of your digital resources. These tests are being worked out at the current time and should be available within a year or so.
- Prepare for accidents. If your counter-disaster plan does not cover digital resources then update it now!

Refreshing data To preserve the information it must be refreshed, a process of copying each bit onto new media to counteract the weakening of the magnetic, optical, or electronic coding, keeping the same order and the same total number of bits.[30] There are various well-established techniques, such as checksums and digests, for tracking the bit-level equivalence of digital information resources and ensuring that a preserved resource is identical to the original.

Refreshing works when the information is encoded in a format that is independent of the particular hardware and software needed to use it, software exists to manipulate the format in current use, present and past versions of software are backwards compatible, and competing hardware and software product lines interoperate.

However, refreshing cannot serve as a general solution for preserving digital information because digital information is produced in highly varying degrees of dependence on particular hardware and software. Most word-processed documents, for example, wrap and embed structure and layout instructions around and in the content. In addition, software is short-lived and market-driven. If the software for making sense of the bits (that is for retrieving, displaying, or printing) is not available, then the information will be, for all practical purposes, lost. In addition, backwards compatibility is both costly and technically difficult for vendors to assure. It is likely that all vendors will maintain a high degree of backwards compatibility with the most recent version of their product to maintain their market share. The rule-of-thumb, however, is that backwards compatibility falls off dramatically after two successive product generations. Finally, interoperability is not only technically difficult for vendors to assure but also fraught with legal restrictions.

Migrating data In the introduction to its final report, the CPA/RLG Task Force records how it 'started from the premise that migration is a broader and richer concept than *refreshing* for identifying the range of options for digital preservation. Although data migration is a common, if difficult, practice as businesses and other organisations preserve their essential business records through successive changes in hardware and business management software, the Task Force regards migration as an essential function of digital archives'.[31]

The following definition is provided: Migration is a set of organised tasks designed to achieve the periodic transfer of digital materials from one hardware/software configuration to another, or from one generation of computer technology to a subsequent generation.[32]

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.[33] Migration includes refreshing as a means of digital preservation but 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 resource as hardware and software change and still maintain the compatibility of the resource with the new generation of technology. In the life cycle of any information resource effective and efficient refreshing and migration take place when its supporting technology peaks. Beyond this point, all preservation action become more expensive as replacement technologies dominate the landscape.

Emulating hardware and software and operating systems This strategy involves emulating obsolete systems on future, unknown, systems, so that a digital document's original software can be run in the future despite being obsolete. In his January 1995 article in *Scientific American*, Jeff Rothenberg suggested that there may be sufficient demand for entrepreneurs to create and archive emulators of software and operating systems that would allow the contents of digital information to be carried forward and used in its original format.[34] While appealing, the suggestion that dependence on hardware and software can be substantially reduced in this way seems to over simplify the magnitude of the problem. Rothenberg also writes in the same article that 'the information revolution derives its momentum precisely from the attraction of new capabilities'.[35] Comprehensive emulation models are not being actively developed in Australia although some limited tests have been carried out at the NLA and European and U.S. trials are being closely watched.

Preserving meaning over time At the far end of the preservation continuum is the challenge of preserving meaning over time. Imagine, for example, digital information resources that have been carefully handled and stored, frequently refreshed to renew their magnetic signals and either migrated to successive technology platforms or kept useable through the use of emulators. Costs and benefits will have been weighed up at each decision point, with the latter exceeding the former. There may however be the need to extract the meaning or essence from resources and re-start the process with whatever storage media and format ensures the most likelihood of further survival. This is the area of automatic document abstracting.

Automatic abstracting operates by taking a source text in electronic form, processing it in order to identify the most important ideas discussed in the text, and organising the selected material to produce a

grammatically acceptable text. Technically this achieved by identifying the most frequently used words or phrases (after discarding definite and indefinite articles, etc). The sentences in which groups of these words occur are then selected and used in the order in which they originally appeared. To try and improve the identification of meaning, weighting can be given to sentences from key locations (e.g. any string which appears in the title of the paper), strings that begin with upper case characters, and to sentences containing cue words (e.g. *finally*, which suggests that a conclusion is starting). Hopefully, these express the essence of the source text in a concise form.

Glenda Browne, in her excellent overview of automatic indexing and abstracting, believes that 'after recent developments in natural language processing by computers, it is now possible for a computer to generate a grammatically correct abstract, in which sentences are modified without loss of meaning.[36]

Digital archaeology If all else fails there is a possibility that digital material may be *rescued* by figuring out how to read stored bits and work out what they mean when systematic archiving has not been performed. Caroline Arms described this strategy as *digital archaeology* in her article in the June 2000 issue of *D-Lib Magazine*.[37]

Printing out to paper and writing out to film In the early 1990s the recommended strategy to preserve digital information was to print it out to paper and preserve it along with other paper-based resources. By the mid 1990s the invention of the World Wide Web prompted a reappraisal of this strategy and the realization that printing out to paper could not be a preservation strategy for *born digital* information resources. Retaining the *look and feel* of these dynamic resources would require some or all of the strategies described at the start of this section. However, printing digital information resources to paper may still have a place in the preservation repertoire for stand-alone documents, particularly if the paper used is long-lasting *permanent paper*.

Computer Output Microfilm (COM) has also been proposed as a long-term preservation *safety net*. COM technology uses an electron beam recorder to *write* a digital photographic image onto highly stable and durable polyester base microfilm; the technology used to create photographic images from satellite and space exploration transmissions. While it is expected that digital information resources will last for many years through migration of data from technology to technology, COM produced in this way approaches perfect microfilm. It exceeds all current technical standards. COM provides even exposure across each frame, zero skew, a QI of 12, a resolution equivalent to 1200 dpi, an image filling the frame, and exact frame placement. It is anticipated that if the digital file eventually deteriorates, re-scanning from the COM will once more produce a very high quality digital file.[38]

Conclusions

The challenge is to ensure that digital information resources remain accessible for as long as they are required. To achieve this we require a fundamentally new approach to identifying, describing, storing and making available all information resources. This approach has to work across multiple resource types, industry sectors, countries and time.

As a democratic society we need significant information to be kept available into the future, irrespective of its analogue or digital nature. A means of ensuring the collective memory has to be found. Similarly the cultural and information sectors cannot afford to work in isolation, either from each other or from their users. It is ineffective and inefficient to do so.

It is also counter-productive to work without cognisance of the major developments throughout the world. We need to build more international linkages that foster common cost-effective strategies and benchmark against the *best of the best*. Technology transfer is the order of the day. In particular, Australia needs to find those areas where it can be the best and others are required to learn from us.

Finally, preservation of digital information resources has to be sustainable. Unless some new business model emerges that leverages public domain information resources into a wealth generating commodities, their preservation has to be paid for from the public purse, which is likely to be static in size or shrink in real terms. This means taking the hardest decision of all; deciding what to preserve from the whole gamut

of old, new and future resources in a never ending re-evaluation of their worth. We need guidelines for what can and should be saved.

This is about 'organising ourselves over time and as a society to manoeuvre effectively in a digital landscape'. It is a problem of building, almost from scratch, the various systematic supports, or 'deep infrastructure', that will enable us to tame anxieties and move our cultural records naturally and confidently into the future.[39]

Notes

---

[1] Global Reach http://www.glreach.com/ [cited 14 July 2000]

[2] G Barker '95% of Children Logging On' *The Age* 22 November 2000 p 9

[3] *Cyveillance: Minding Your Business on the Net* http://www.cyveillance.com/newsroom/pressr/000710.asp [cited 10 July 2000 [cited: 29 July 2000].

[4] B Kahle 'Preserving the Internet' [online] *Scientific American* March 1997 http://www.sciam.com/0397issue/0397kahle.html [cited 24 July 2000]

[5] P Lyman and B Kahle 'Archiving Digital Cultural Artifacts: Organizing an Agenda for Action' [online] *D-Lib Magazine* July/August 1998 http://www.dlib.org/dlib/july98/07lyman.html [cited 6 August 2000]

[6] T Kuny 'The Digital Dark Ages?: Challenges in the Preservation of Electronic Information' *International Preservation News* no 17 May 1998 pp 8-13 http://www.ifla.org/VI/4/news/17-98.htm#2 [cited 24 July 2000]

[7] J Rothenberg 'Ensuring the Longevity of Digital Documents' *Scientific American* January 1995 p 29

[8] S Gilheany 'Preserving Information Forever and a Call for Emulators' Presented at *Digital Libraries Asia 98: The Digital Era: Implications, Challenges & Issues* 17-20 March, 1998 URL http://www.archivebuilders.com/ [cited 5 December 2000]

[9] J Van Bogart *Magnetic Tape Storage and Handling: A Guide for Libraries and Archives* Washington DC The Commission on Preservation and Access and National Media Laboratory 1995 http://www.clir.org/cpa/reports/pub54/acknowledgements.html [cited 6 December 2000]

[10] Rothenberg 1995 op cit

[11] Kodak 'So, How Long Can CDs last?' http://www.kodak.com/cluster/global/en/professional/products/storage/pcd/techInfo/permanence.shtml

[12] J Van Bogart 1995 op cit

[13] R Harvey 'The Longevity of Electronic Media: from Electronic Artefact to Electronic Object' *Multimedia Preservation: Capturing The Rainbow, proceedings of the second national conference of the National Preservation Office Brisbane 28-30 November 1995* Canberra ACT National Library of Australia 1995 pp 202-216 http://www.nla.gov.au/niac/meetings/npo95rh.html

[14] T Kuny 1998 op cit p 8

[15] T Kuny 1998 op cit p 9

[16] R Harvey 1995 p 202

[17] Kulturarw^3 http://kulturarw3.kb.se/html/kulturarw3.eng.html

[18] Eva http://renki.lib.helsinki.fi/eva/english.html

[19] B Kahle 'Preserving the Internet' [online] *Scientific American* March 1997 http://www.sciam.com/0397issue/0397kahle.html [cited: 24 July 2000]

[20] T Kuny 1998 op cit p 9

[21] H Berthon and C Webb 'The Moving Frontier: Archiving, Preservation and Tomorrow's Digital Heritage' [online] Paper presented at *VALA 2000 – 10th VALA Biennial Conference and Exhibition* Melbourne Victoria 16 – 18 February 2000 http://www.nla.gov.au/nla/staffpaper/hberthon2.html [cited 14 March 2000]

[22] J Garrett and D Waters *Preserving Digital Information: Final Report and Recommendations 20 May 1996* Washington DC Commission on Preservation and Access Task Force on Archiving of Digital Information http://www.rlg.org/ArchTF/ [cited: 21 April 2000]

[23] J Garrett and D Waters 1996 op cit p iii

[24] J Garrett and D Waters 1996 op cit p 40

[25] A Howell 'http://www.na.gov.au/padi/: Preserving Access to Digital Information (PADI) – an Opportunity for Global Cooperation' Paper presented at *Managing the Preservation of Periodicals and Newspapers* Paris 20-24 August http://www.ifla.org/VI/4/conf/howell.pdf [last revised 14 August 2000]

[26] National Library of Australia *PADI Metadata Schema* http://www.nla.gov.au/padi/metadata.html

[27] State Library of Tasmania *Our Digital Island* http://www.tased.edu.au/library/odi/index.htm

[28] Commonwealth of Australia *National Conservation and Preservation Policy and Strategy* Department of Communications and the Arts Canberra ACT 1998

[29] M Lesk 1996 'Preserving Digital Objects: Recurrent Needs and Challenges' *Multimedia Preservation: Capturing The Rainbow, Proceedings of the Second National Conference of the NationalPreservation Office Brisbane 28-30 November 1995 Canberra ACT National Library of Australia* p 106

[30] M Lesk *ImageFormats For Preservation and Access: A Report of the Technology Assessment Advisory Committee to the Commission on Preservation and Access* Commission on Preservation and Access Washington DC 1990 p 5

[31] J Garrett and D Waters 1996 op cit p iii

[32] Ibid

[33] Ibid

[34] Rothenberg 1995 op cit p

[35] Rothenberg 1995 op cit p

[36] G Browne 'Automatic Indexing' *LASIE* vol 27 no 3 pp 58-65

[37] C Arms 'Keeping Memory Alive: Practices for Preserving Digital Content at the National Digital Library Program at the Library of Congress' [online] *RLG DigiNews* 15 June 2000 http://www.rlg.org/preserv/diginews/diginews4-3.html [cited 24 July 2000]

[38] A Howell 'Digital Imaging Technology for Preservation and Access: A Cornell University Library Workshop' *LASIE* vol 27 no 1 1996 pp 26-41

[39] J Garrett and D Waters 1996 op cit p iii