The moving frontier: archiving, preservation and tomorrow's digital heritage

Hilary Berthon
Manager
National & International Preservation Activities
National Library of Australia
hberthon@nla.gov.au

Colin Webb
Director
Preservation Services Branch
National Library of Australia
cwebb@nla.gov.au

Abstract:
Digital publications are a significant part of tomorrow's heritage of digital information. However, there is a growing understanding that tomorrow's digital heritage will simply not be available without concerted action. This paper reviews international progress in digital archiving and preservation over the past one to two years. In that time, we have seen some developments in international collaboration, many archiving models being tested, active work on a range of facilitating issues, and an ongoing debate over the most appropriate long-term preservation strategies. However, a number of problematical issues remain. A most encouraging trend is the ongoing commitment to sharing information. The National Library of Australia's PADI website has been re-developed as an international digital preservation forum, charting progress in finding workable solutions that can be applied by Australian libraries.

Tomorrow's digital heritage is very largely a product of our own era. The fact that we use such a term as "digital heritage" suggests two important things: that there are parts of it we want to keep, and that we recognise it is fragile and at risk. The purpose of this paper is to look at recent progress in keeping our digital heritage accessible.

The development of traditional library preservation has involved a number of steps: recognising problems needing attention, developing responses, sharing information and establishing orthodoxies from both conceptual and heuristic investigation. Innovation has been fostered by factors such as: economic pressures, changing attitudes to the ethics of intervention, changing expectations about what is meant by "preserving" or "conserving", and the appearance of enabling technologies. We are now seeing similar serious work and forces for innovation in preserving access to different kinds of digital collections.

The National Library of Australia is one of many players looking for feasible ways to maintain access to our digital heritage. Just as it is important to participate in finding solutions to our own needs, we think it is also important to discover and critically consider the work and ideas emanating from around the world, and to share these with others in the Australian library community. It was with this perspective that the National Library created the PADI website[1] in collaboration with a number of other Australian institutions, and assumed responsibility for its maintenance and development.

We would like to regularly review progress in digital preservation, especially as it relates to the Australian library sector, based on our view through the PADI window. We would like to identify the main issues being addressed, the main groups who are involved, and the main directions being explored. We will focus on some areas that look to be most promising, or where there has been most activity that we are aware of. These include: the development of models for collaboration; models for archiving; and approaches which may assist in preservation, such as persistent identifiers and metadata schemes to support the long-term management of digital objects. We will also discuss the current attention to formats that may facilitate preservation, and the development of strategies for dealing with hardware and software dependencies. In looking at all of these we find ourselves emphasising some recurring threads: the need for testing of concepts and theories as concretely and specifically as possible, the continuing difficulties of dealing with complex multimedia publications, and the importance of encouraging communication.

We are still in the early stages of developing ways of maintaining access to digital material. There is still a great and legitimate concern about our ability even to put workable archiving arrangements in place to provide the first steps in ensuring long-term accessibility. It is not at all surprising that most effort has focused on this part of the process, although as one reviewer put it: "… [the processes] which address long-term preservation itself receive little attention, even though over the long term they will consume far more resources and management time. This is as if digital preservation were an iceberg, and the cultural world is concentrating uniquely on the visible parts, ignoring the much bigger problems below the water line…. This focus to immediate 'above the water line' issues is understandable: it addresses immediate issues, and these early practices are essential first steps towards long-term preservation. However, it is important that longer term issues are also addressed."[2]

Sharing ideas: models for international collaboration

At the broadest level, many institutions have common interests, and initiatives for international consultation and cooperation have been welcomed.

In 1998, the National Library of Australia established an international digital collaboration group with eight other national libraries and three consortia, all known to be active in digital preservation.[3] We have used this collaboration to consult and share ideas on a number of the issues referred to in this paper. Despite the tension between finding time to consult and time to progress institutional programs, this and other cooperative initiatives have been adopted enthusiastically.

Models for national and regional archiving collaborations

In its seminal report issued in 1996, the (US) Taskforce on Archiving of Digital Information recommended the establishment of a national system of certified digital archives. These archives would be collectively responsible for maintaining long-term accessibility. The report argued that the most effective and affordable strategy in a time of immense change would be a distributed system for collecting, protecting and preserving digital information. Such a strategy would assign responsibility to "those who presumably care most about and have the greatest understanding of the value of particular digital information objects."[4]

We are seeing a range of organisational models evolving in different countries, and as noted in more recent reports, a variety of players are poised to play a role in digital preservation, including legal deposit libraries, data centres, digitisers, universities, research institutions and publishers.[5]

In the USA, the Research Libraries Group (RLG) and the Digital Library Federation (DLF) have set up a joint taskforce to look at policy and practice for long-term retention of digital material. The group includes an impressive array of expertise and a worthwhile brief: to "gather and analyse existing digital preservation policies and practice descriptions for the following classes of electronic materials: institutional records in digital form (ie electronic records); locally digitised materials (institutional projects); and electronic publications." It will "create one or more digital preservation policy frameworks that adequately address the material types and relevant institutional contexts …(and) will make the policy framework(s) widely available…"[6] The task force is due to report in March 2000.

In Canada, a national archiving model is yet to emerge, but the National Library of Canada (NLC) has proposed several possible cooperative arrangements, in a policy document issued in October 1998.[7] The options included: a distributed model possibly including multiple copies of publications stored at different locations; a distributed access but centralised preservation model in which the NLC might store electronic hard copies of publications also stored by other parties on remote servers; or a centralised model in which the NLC would work with other federal agencies to create a central repository for electronic information.

The Canadian Initiative on Digital Libraries (CIDL) consortium may also play a part. With a membership of more than fifty Canadian libraries in the academic, public and special sectors, CIDL aims to "promote, coordinate and facilitate the development of Canadian digital collections and services in order to optimise national interoperability and long-term access to Canadian digital library resources."[8] Part of its brief is to look at the issue of assigning roles and responsibilities for long-term archiving of digital material.

The European Commission has been a very active supporter of regional collaboration in Europe. Among many interesting projects supporting access to digital heritage[9] is the NEDLIB project, a collaborative project involving eight European national libraries, one national archive, two IT development companies and three major publishers, coordinated by the National Library of the Netherlands. The project "aims to construct the basic infrastructure upon which a networked European deposit library can be built"[10] in order to ensure ongoing access to electronic publications.

In the UK, a number of models have been successfully adopted for different types of digital material. The Arts and Humanities Data Service (AHDS)[11] is a centrally managed but geographically distributed service which has assumed a significant role in preserving access to digital resources in the arts and humanities field for the higher education sector. It plays an active role in educating the scholarly community about matters such as the impact of decisions made at the point of a digital object's creation on its long-term accessibility. AHDS also actively identifies and promotes the use of standards to facilitate future access.[12] The Natural Environment Research Council data centres are also based on a physically distributed model in which materials are housed at locations where there is expertise to manage them.[13] Perhaps reflecting the success of this approach, a recent UK study, funded through the higher education sector's Joint Information Systems Committee, [14] recommended that a body should be established to coordinate archiving of digital materials in the UK but that the job of maintaining the archives should be contracted out to specialist agencies with the appropriate expertise. It was proposed that the coordinating body might be formed as an extension of the National Preservation Office (NPO), currently supported by the British Library with additional financial support from the Public Record Office, The Consortium of University Research Libraries, and five major UK libraries.

In Australia we have been developing a national collection of electronic publications built on shared responsibility for archiving by the National and State libraries. This is still evolving: while a number of State libraries (Victoria and South Australia) have decided to be part of the PANDORA[15] architecture developed by the National Library, some other State libraries (in particular Tasmania and New South Wales) are pursuing their own systems. We particularly value the evolution of this loose national network and its anticipated gradual extension to other significant stakeholders such as university libraries and major publishers with an interest in archiving. We value the opportunity to develop a national model, but we also value the chance to work with a diversity of approaches. The archiving of electronic theses being pursued by university libraries through the Australian Digital Theses Project[16] is an excellent illustration of the way a national archiving and preservation model is likely to develop: as a patchwork of action by players with particular business needs and interests.

Models for archiving

As well as the collaborative arrangements which have been established, we have seen a range of archiving models under development. These models are emerging both internationally and through a number of national and local level projects. There is no consensus yet on the "best" archiving model, but this is not surprising given the diversity of user communities with archiving needs. We may never have a single archiving model except at the broadest level.

Internationally, increasing attention is being given to the Reference Model for an Open Archival Information System (OAIS) being developed by the Consultative Committee for Space Data Systems as a new ISO standard.[17] This model provides terms of reference, conceptual data models, and functional models for open archives that can interoperate. It defines the nature of "information packages" in terms of both the content and what is needed to understand, access and manage the content. The model also aims to describe the processes required for archiving to be successful. Clearly this is a very helpful model, of interest to everyone concerned with archiving digital information. A number of archiving projects, including the NEDLIB and CEDARS[18] projects, are trying to follow the OAIS model closely. Others, such as our PANDORA project, are using its concepts to inform the architecture they are developing. This approach sees the Reference Model more as a checklist of requirements to be addressed than as a definite road map for the archive and all its structures and processes.

At the national and regional levels a number of archiving projects are hammering out frameworks and implementation models reflecting their own needs. Probably all archiving endeavours dealing with digital library information can still be considered experimental to some degree: most are still under development, operating at low volumes or dealing with a subset of expected problems.

Of immense importance in the development of archiving models is the 3-year CEDARS project, being undertaken under the overall direction of the (UK) Consortium of University Research Libraries (CURL), and with funding from JISC, which aims to address the "strategic, methodological and practical issues and will provide guidance for libraries in best practice for digital preservation."[19] The archive architecture developed by the CEDARS project is an implementation of the OAIS model for a distributed digital archive of library resources.[20]

As already mentioned, the NEDLIB project has also adopted the OAIS model as a framework on which to build its archiving model for deposit libraries. It has added a Preservation Module that includes provision for both emulation approaches and digital migrations resulting in changed content.[21]

Both the British Library and the National Library of Australia have been actively developing specifications for digital management systems that will include archiving as one of several key functions.

The British Library (BL) released a Briefing Document in July 1999, providing information on the approach they are taking with their Digital Library Programme. This approach has been to purchase a digital library system (DLS) as technical

infrastructure to serve as the basis of the UK national digital archive, enabling the BL "to store, preserve and provide access to the UK digital published output".[22] One requirement is that the overall functional design should conform to the outline model of the OAIS Reference Model. Another is that is must facilitate interoperability with other similar developments in the UK and elsewhere, in particular CEDARS, NEDLIB, AHDS, and the National Library of Australia's Digital Services Project (DSP).

The NLA's DSP[23] grew out of our PANDORA proof-of-concept archive for Australian online publications, and the need to improve our management of a range of other digital and non-digital collections. It aims to exploit the opportunity presented by digital technology to integrate at least some levels of discovery, access and management of all of the Library's collections. The DSP has evolved into a metadata repository and powerful search facility, for which tenders were let in late 1999, and a digital collection management system, for which tenders are currently being evaluated. While taking account of all the archiving models known to us, the DSP specification has developed out of our own particular requirements to manage the processes of archiving, access provision and long-term management. It has developed as a robust model partly because we have been actively involved in archiving digital publications for some years, and partly because we have invested heavily in a strong and evolving policy framework on which our archiving activity has been built. PANDORA remains selective in its collecting approach, and aims to ensure that what is collected is made accessible and remains accessible. The selective approach makes it possible to apply bibliographic control to the archived files, and to check the effectiveness of the archiving processes applied to each file.

The Swedish Kulturarw[3] project[24] takes quite a different approach. While it also aims to test methods of collecting, preserving and providing access to online electronic documents within its own national domain, it attempts to download everything, rather than undertake a detailed selection process. This approach offers the prospect of more completely automated collecting processes, and a more comprehensive archive of national online publishing activity, if the archive's accessibility objectives can also be achieved. The Finnish EVA project has adopted similar methods for acquiring a national collection of online publications.[25, 26]

The NLC's Canadian Electronic Publications Pilot Project (EPPP), commenced in 1994, aimed "to identify and understand all the challenges associated with acquiring, cataloguing, preserving and providing access to Canadian electronic publications."[27] A small number of Canadian electronic journals and other representative publications freely available on the Internet were used for the pilot project. The EPPP was used to develop an archiving model that provides the basis of current routine operations at the NLC.[28] The Electronic Collection,[29] initiated in 1997, by mid-1999 had developed a comprehensive collection of government publications and a significant if comparatively small number of publications issued by commercial publishers.

There are many other major libraries just beginning to actively investigate digital archiving including the National Diet Library in Japan and the Swiss National Library.[30]

Of course, archiving of digital data is not new, nor is it the exclusive domain of libraries and projects dealing with "published" information. As already implied by the fact that the OAIS Reference Model is being developed by the space data community, there are many communities dealing with the need to preserve many different kinds of data. The Arts and Humanities Data Service, already mentioned in the context of national and regional collaborations, encompasses a number of service providers specialising in particular kinds of data, such as archaeology and mapping data, text, visual arts data, and performing arts data. These service providers are required to develop archiving programs appropriate to their collections and their user communities, within the overarching policy framework set by AHDS.

We have much to learn from the way such data archiving communities deal with their problems. However, in keeping with the spirit of learning from diversity, our task is always to recognise what is really relevant only to a particular set of archiving needs, and what may be more generally applicable.

Approaches to assist in preserving access

Persistent identifiers

As a component of many of these archiving models, and in wider discussion, there has been interest in a range of approaches designed to facilitate long-term access.

One such focus has been to find ways of identifying digital objects available on the Internet, so that they remain findable. Although URLs serve to identify resources and describe their location on the World Wide Web, they are notoriously unreliable as they must change whenever a digital resource moves to a new location. Recent attempts to develop more persistent identifiers are driven by the need to provide reliable ongoing access regardless of location on the Internet.

The Uniform Resource Identifier (URI) architecture aims to be a comprehensive and persistent resource discovery system for the Internet. This architecture consists of: Uniform Resource Names (URNs) – standard, persistent and unique identifiers for digital resources on the Internet; URLs for locations; and Uniform Resource Characteristics (URCs) – standardised metadata about the resource. Finally, for a user to be able to link to the URL of a digital resource from the URN, a resolver service is required.

A standard URI-architecture has not been realised. While the Uniform Resource Names Working Group of the Internet Engineering Task Force[31] has been developing standards for the URN, a number of unique identifier systems have already been implemented by different groups. These include The Handle System[32], the Digital Object Identifier (DOI) initiative[33] and the Persistent URL (PURL).[34]

A meeting of the Conference of Directors of National Libraries in 1998 agreed to set up a Persistent Identifiers Task Force, chaired by Winston Tabb of the Library of Congress. The Task Force met in Washington in April 1999, and concluded that while persistent identifiers (PIs) are critical to preserving access to digital information, no completely satisfactory system has emerged that all parties could undertake to use.

Diverse needs and approaches mean that a plurality of identifier systems is likely to evolve (and is evolving), so their interoperability is crucial.

At the NLA, we have recognised that we are using something like *de facto* PIs whenever we name a digital file and undertake to maintain external access to it. Having implemented a number of file naming conventions for different digital collections over recent years, we have decided to look at how these naming systems might serve as PIs. We will examine their potential benefits, associated costs and ability to assist with archiving and preservation.

At the same time as we are using our existing collections as a testbed, the State Library of Tasmania has been funded to develop a PI system for digital resources for the Tasmanian government sector. We expect these and other archiving projects to take us towards a workable PI system, or network of systems, in Australia over the next few years.

Preservation metadata

Another approach to ensuring long-term accessibility has been the development of structured ways of describing the preservation management requirements of digital resources. This preservation metadata may be used to store technical information that supports preservation decisions and action, to document preservation action taken such as migration or emulation, to record the effects of preservation strategies, to ensure the authenticity of digital resources over time, and to note information about collection management and the management of rights. In contrast to descriptive metadata schemas (eg MARC, Dublin Core), which are used in the discovery and identification of digital objects, preservation metadata largely falls into the category of administrative metadata, assisting in the management of information.[35]

The development of preservation metadata has been attempted in a number of recent projects. The resource intensive nature of digitisation projects has prompted a number of groups to develop preservation metadata to ensure that the digital content created can be maintained.[36] However, these metadata sets are often inadequate for "born digital" items. Both the CEDARS and NEDLIB projects have been active in developing metadata sets to manage collections.[37]

At the National Library of Australia, work in this area has been prompted by the need to define what information we will need in order to manage long-term preservation of our digital resources.[38] We believe we will only be able to manage our growing archive of digital files, in a great range of formats, through the intelligent use of metadata. It is impossible to determine unequivocally what we will need to know in order to manage digital preservation in the future, so our metadata elements necessarily reflects assumptions about our future requirements. We also recognise that different types of digital materials, and different archiving systems, will need different metadata support. We have tried to specify the information we need from a system in order to support the decisions we will have to make, rather than attempt to prescribe what data should be entered at particular stages and by whom. This is a metadata-output model that we believe should be applicable to many implementations that may decide to record this information in a variety of ways.

Standards

Generally speaking, there has been a move away from relying solely on standards to solve digital preservation problems. It has been recognised that the diversity of needs, the pressure for change which limits market support for long-term standards, and the difficulty in predicting future changes in technologies, all mean that we are likely to continue to see a plethora of standards of short-term applicability. Although the development of standards has been critical to interoperability and to automating processes, standards are likely to play a role in facilitating preservation rather than being a complete solution.

A major focus in recent years has been the use of the standard structured information formats: SGML and its simplified form, XML. The potential for sorting collection items into categories of similar materials for batch preservation action is attractive, as is the encapsulation effect of mark-up languages that separates the content from the structure and formatting. While this is a big step forward – and XML promises to be applicable to many different format types – marking up documents to comply with standards is highly resource intensive. It is also inevitable that libraries will continue to deal with digital objects that are not fully compliant with standards.

Essence

Finally, preservation will be helped by attempts to define the "essence" that must be preserved, and procedures for authenticating its survival over time. This is a complex issue in which developments are likely to emerge as we encounter the realities of the preservation strategies we choose to use and their effects on the objects we are seeking to preserve. Important conceptual work has recently been carried out[39] on developing a canonical reference point for each digital object, so that later versions can be compared with the "essence" that someone has decided should be preserved. It will be interesting to see how this develops, given the resources required, and the compromises in object integrity likely to be a part of any preservation strategy.

Further archiving issues

Further archiving issues which we have identified as presenting particular challenges include:

- collecting, managing and providing access to dynamic, software-driven objects like databases. While these types of objects present challenges for long-term management, we have found it difficult to even bring them into an archive.
- the scalability of our archiving models. This issue was identified in a European Commission sponsored study[40] 12 months ago and it remains a concern. Projects such as PANDORA, NEDLIB and others will be crucial in providing us with information on the issues in managing the increasing numbers of items and large amounts of data expected to come into collections over the long term.
- the development of relationships between archiving institutions and publishers, principally over intellectual property concerns. It is understandable that there are unresolved tensions in a period of such profound change. There is a widely held and emerging perspective that these tensions can be worked through by keeping

the interests of all stakeholders in mind and looking for arrangements that maximise mutual benefits.
- the need for resources to undertake even relatively simple things like identifying digital items that have entered collections and assessing their current accessibility. For many libraries even these first steps are hard to achieve.

Formats for Preservation

We have seen an ongoing interest in developing formats that will be preservable. Most of these attempts involve some kind of encapsulation, or grouping together, of a digital object and anything else needed to provide access to it. Encapsulation can be achieved by using physical or logical structures called "containers" or "wrappers" to provide a relationship between all information components, such as the digital object and supporting information including unique identifiers, metadata, and software specifications. Encapsulation may also encompass the software itself required to read the object. The "package" may be composed of analogue and digital components. An example of an analogue component would be human readable instructions, such as writing on the outer case of a physical format carrier, describing how to use the carrier and interpret the outermost layer of the digital component, most likely the wrapper, which will in turn provide the information required to use the rest of the digital information contained. An alternative to storing all the supporting information with the object is to include a clear pointer to a single, reliable storage area for that information.

Obviously, the encapsulation concept is used in many CD-ROM publications that come with their own operating software and instructions for use. However, preservation interest in encapsulation is based on encapsulating enough information and/or software to allow the digital object to be useable across changes in operating systems over time.

This concept underlies the OAIS Reference Model which employs the concepts of "information packages" (IPs) that are composed of "content information" and "preservation description information" contained by "packaging information". The content information, in turn, includes the actual digital object and the "representation information" needed to interpret it. The preservation description information portion includes information about provenance and context, reference information (such as unique identifiers) and a wrapper which protects the object against undocumented alteration.

Encapsulation has also been explored in a range of other projects and papers. The Universal Preservation Format[41] project has focused on developing a "self-describing" platform-independent format which includes, within its metadata, all the technical specifications required to build and rebuild appropriate media browsers to access contained materials throughout time. Another proposed model describes the encapsulation of hardware, software and data in the form of a "digital tablet".[42] The Digital Rosetta Stone[43] concept relies on keeping a 'meta-knowledge archive' of how to interpret media formats and file formats to support data recovery and document reconstruction processes. For efficiency, it is proposed that this representation information be stored separately from the encapsulation.

Very long-lasting media, in conjunction with encapsulation or secured metadata, have been suggested as a way of allowing documents to remain useable for very long periods of time. This model underlies both the HD-Rosetta product developed by Norsam Technologies under license from the Los Alamos National Laboratory in the US[44], and the digital tablet (*vide supra*).

These approaches all hold some potential, but are more likely to be helpful in informing other approaches than in being widely adopted themselves. For example, while some may prove to be prohibitively expensive for widespread adoption, they do help us think more about ways of associating content with the tools needed to understand and use it, and recognise that media instability will continue to be an issue wherever we try to store digital information for long periods.

Dealing with hardware and software dependencies

While media deterioration causes some concern, it is the prospect of relentless changes in information technologies that threaten to render digital materials rapidly inaccessible – "held hostage to their own encoding".[45] In recent years a range of approaches to overcoming technological obsolescence has been proposed and debated, and this debate has continued over the past year. In January 1999, the (US) Council on Library and Information Resources (CLIR) published a report[46] by Jeff Rothenberg of the Rand Corporation, which discussed various proposed solutions to long-term digital preservation and elaborated an emulation strategy proposed by himself and others in recent years.

Emulation refers to the process of mimicking in software a piece of hardware or software so that other processes act as if the original equipment or function is still available in its original form. The emulation strategy described by Rothenberg entails emulating obsolete systems so that the digital object's original software can be run. In contrast to migration, in which the original object is successively transferred to new systems, once the data is archived with appropriate metadata and software this emulation model requires no action apart from media refreshing and transfer until access is desired, at which time an appropriate emulator is either found or developed.

Although emulation is a proven technique in current computer systems, there is widespread agreement that its feasibility and benefits are still to be proven for preserving access to large numbers of complex digital objects. Practical studies are planned as part of a collaborative effort involving the CEDARS project team and researchers at the University of Michigan with funding through the Joint NSF/JISC International Digital Libraries Initiative.[47] Their project aims to "develop and test a suite of emulation tools, evaluate the costs and benefits of emulation as a preservation strategy for complex multi-media documents and objects, and develop models for collection management decisions about how much effort and resources to invest in exact replication within preservation activity… (It) will assess options for preserving the original functionality and 'look and feel' of digital objects and develop preliminary guidelines for the use of different preservation strategies (conversion, migration and emulation)."[48]

The National Library of the Netherlands will also undertake a project to test the emulation strategy for long-term preservation. This project, to be carried out in

collaboration with Jeff Rothenberg, will test the viability of using hardware emulation as a means of preserving digital publications in a deposit library. The testbed environment of the Deposit System for Electronic Publications (DSEP) developed in the NEDLIB project, using sample material provided by NEDLIB-sponsoring publishers, will be used to carry out experiments for this project.[49] Both these projects will be crucial in discovering the potential of this largely untested approach to digital preservation.

The recent paper by Rothenberg added heat to the long-running debate between proponents of emulation and migration strategies. Rothenberg attacked the migration strategy as requiring "continual heroic effort" that would, in any case, fail to maintain document integrity. Rothenberg's approach, and the emulation strategy, in turn received criticism from a number of sources, including David Bearman, who warned of the "potentially dangerous wishful thinking" which would result in unquestioning reliance on the ability of emulators to later access digital material and asserted that "Rothenberg is fundamentally trying to preserve the wrong thing by preserving information systems functionality rather than records".[50] Bearman does, however, see common ground in the specification of metadata which is needed to support all digital preservation strategies.

Although emulation has received much attention, migration still looks like the strategy of first choice. In the US, the Digital Library Federation recently published a detailed case study of migration in a specific environment: a social science data archive.[51] The report goes well beyond conceptualising to analyse concrete experience in a controlled but real business environment. The data being migrated is largely homogenous, and yet the report reveals many difficulties, and many useful lessons. In particular, it emphasises the importance of maintaining the documentation necessary to understand the data.

One approach which continues to receive bad press is technology preservation – the maintenance of machines and software in the hope that they will continue to provide access to the digital objects that used them. While this looks no less futile in the long term than it ever did, we need to admit that it remains the key bridging strategy for most of us, whether it is simply trying to hold onto old machines that still work, or maintaining software archives to run the digital objects in our collections.

Finally, the recovery or rescue of digital information from media or formats that have become inaccessible has received some attention. The best review can be found in one of the seven very interesting and useful JISC/NPO studies on the preservation of electronic materials published between 1997 and 1999.[52] This paper discusses a number of problems and scenarios, and reports on programs to rebuild computers, to build simulators and emulators, and data recovery required as a result of disasters or poor management of technological change. Although most data can be recovered given sufficient resources, the time and money required are such that a strong case needs to be developed to show that the data is worth the cost of recovery. (It also means that "data archaeology" is an expensive and unreliable substitute for more anticipatory preservation activity.)

It is not yet clear whether there will be a single preservation strategy which emerges as being best for all types of digital library material. Given the diverse collections for

which libraries are responsible, it appears most likely that combinations of strategies will be used in the foreseeable future. At the National Library of Australia, for example, we know we need:

- some data recovery for the large number of old and poorly documented floppy disks that are currently inaccessible: we are experimenting with some data recovery software to establish guidelines and procedures for this;
- media transfer and refreshing for the data we can access that is currently on unstable carriers;
- some technology preservation, including maintenance of software and even some hardware;
- migration strategies to produce versions for current access using contemporary platforms, while attempting to maintain archival copies of versions of files in their native format. This will enable them to be operated using maintained obsolete systems if available, or accessed by emulators if available. It should also enable them to used as the starting point for re-migration if the ongoing migration stream breaks down for some reason.

We will also need to develop and refine our current investigations of matters such as the implications of new versions of HTML for digital preservation, the conversion of complex files formats to alternative storage formats that might be more easily migrated, and our development of a preservation plan for each type of file format. We are aware that whatever strategies we use, we will need excellent technical and management systems, and excellent metadata to support them.

Making progress

We are at an interesting stage in the development of solutions to the evolving challenge of digital preservation and it will be exciting to watch where these developments lead. There are, however, some issues that are either receiving a low level of attention, or are poorly documented. These include:

- cost effective ways of predicting deterioration rates of physical carriers;
- useable indicators of technology change that should guide or trigger preservation action;
- decision models for selecting which preservation strategies are most applicable for particular kinds of complex digital information;
- ways of testing our ability to preserve.

Information sharing and communication

There is a large and growing “literature” (if that is the right word to use), and numerous active communication channels. From this field we have chosen one small example.

One of the most prominent collaborations in the field over the past few years has occurred in the higher education and research sectors in the UK. The eLib program has provided both funds and vision for the investigation of digital preservation issues. The program funds the CEDARS project as a practical way of exploring many of the

issues raised in the seven supporting studies on preservation already referred to. That the studies were guided by a specially formed Digital Archiving Working Group, comprising representatives from the British Library, the National Preservation Office (NPO), the Publishers' Association, universities, data archives, and the Public Record Office (PRO), is indicative of a high level of collaboration focused on systematic research and information sharing. The seven studies are of interest in themselves, but their value was considerably enhanced this year by the publication of a synthesis of the studies.[53] The booklet reflects the value (and many of the shortcomings) of the studies, and has useful information on cost modelling and on some organisational and decision issues. Further indicating its collaborative nature, it was launched at a seminar organised by the NPO, the Library and Information Commission, the Museums and Galleries Commission and the PRO.

PADI and information sharing

Finally, we would like to say more about PADI, our primary vehicle for documenting and sharing information on this critically important subject.

In response to a growing recognition of the need to safeguard digital heritage, the National Library of Australia set up its PADI website in 1997. The Library has since developed this site into a comprehensive subject gateway to resources dealing with a wide range of aspects of digital preservation. The PADI resource plays a role in educating users about the subject areas accessible through the gateway. While users have the option of proceeding directly to a search, a high level of support is available to assist users through the provision of explanatory text. Digital preservation topics dealt with on the PADI site are as diverse as: persistent identification of networked material, the implementation of legal deposit for electronic publications, migration of digital information and preserving access to physical format digital material.

Powerful search capabilities are supported by the assignment of metadata, based on the AGLS and Dublin Core, to linked resources. A thesaurus of digital preservation terms has been developed to provide subject metadata enabling precise retrieval of resources from the PADI database. Certain types of resources may selectively be retrieved: for example, policies, strategies and guidelines, information about digital preservation projects, or articles may be independently selected.

Cooperation has played a crucial role in the PADI initiative which has been carried out in partnership with Australian and international experts. A news and online forum area provides an opportunity for all who are interested in preserving access to digital material to send and receive news about digital preservation and to participate in discussion of digital preservation issues. Users are encouraged to suggest resources for the PADI database using a simple online form. In these ways, PADI provides a focus for information sharing and cooperation supporting digital preservation activities both in Australia and worldwide.

(The authors would like to acknowledge the assistance of Deborah Woodyard of the National Library of Australia Digital Preservation Unit in preparing some of the material for the PADI website used in this paper.)

[1] National Library of Australia. *PADI: Preserving Access to Digital Information*. Online. Available: http://www.nla.gov.au/padi/. 8 October 1999.
[2] M. Fresko, 1998, "Results of the comparative study of digital preservation guidelines", in *Digitisation of Library Materials, Concertation meeting and workshop, Luxembourg, December 1998*
[3] The national libraries are: The Library of Congress, The British Library and the national libraries of Sweden, Finland, Germany, The Netherlands, Canada, and Norway. The consortia are: the Research Libraries Group (RLG), the Digital Libraries Federation (DLF) based in the USA, and the CEDARS Project (CURL Exemplars in Digital Archives) based in the UK.
[4] Task Force on Archiving of Digital Information, Garrett, John, and Waters, Donald (chairs). *Preserving digital information: final report and recommendations.* Commission on Preservation and Access (CPA) and Research Libraries Group (RLG), 20 May 1996. Online. Available: http://www.rlg.org/ArchTF/. 8 October 1999.
[5] Beagrie, N. and Greenstein, D*. A Strategic Policy Framework for Creating and Preserving Digital Collections.* Version 4.0 (Final Draft). Arts and Humanities Data Service, July 1998. Online. Available: http://ahds.ac.uk/manage/framework.htm. 7 October 1999.
Hodge, G., and Carroll, B. C*. Digital Electronic Archiving: the State of the Art and the State of the Practice.* International Council for Scientific and Technical Information and CENDI, April 1999. Available: http://www.icsti.org/icsti/99ga/digarch99_TOCP.pdf, http://www.icsti.org/icsti/99ga/digarch99_ExecP.pdf and http://www.icsti.org/icsti/99ga/digarch99_MainP.pdf. 7 October 1999.
[6] Research Libraries Group, 16 July 1999*. RLG-DLF Task Force on Policy & Practice for Long-term Retention of Digital Materials.* Online. Available: http://www.rlg.org./preserv/digrlgdlf99.html. 7 October 1999.
[7] Electronic Collections Coordinating Group National Library of Canada*. Networked Electronic Publications Policy and Guidelines.* National Library of Canada, October 1998. Online. Available: http://www.nlc-bnc.ca/pubs/irm/eneppg.htm. 8 October 1999.
[8] National Library of Canada. *Canadian Initiative on Digital Libraries* (home page). Online. Available: http://www.nlc-bnc.ca/cidl/. 7 October 1999.
[9] eg BIBLINK (home page: http://hosted.ukoln.ac.uk/biblink/); DESIRE (home page: http://www.ub2.lu.se/desire/).
[10] Koninklijke Bibliotheek. *NEDLIB Networked European Deposit Library* (home page). Online. Available: http://www.konbib.nl/nedlib/. 7 October 1999.
[11] *Arts and Humanities Data Service* (home page). Online. Available: http://ahds.ac.uk/. 7 October 1999.
[12] Beagrie, N., and Greenstein, D*. Managing Digital Collections: AHDS Policies, Standards and Practices.* Consultation Draft, version 1. Arts and Humanities Data Service, December 1998. Online. Available: http://ahds.ac.uk/public/srg.html. 7 October 1999.
[13] Natural Environment Research Council. Version 2.1, March 1999. *The NERC Data Policy Document*. Online. Available: http://www.nerc.ac.uk/environmental-data/data/background.htm . 8 October 1999
[14] Haynes, David, Streatfield, David, Jowett, Tanya, and Blake, Monica. *Responsibility for digital archiving and long term access to digital data.* British Library Research and Innovation Report 67, 1997. Available: http://www.ukoln.ac.uk/services/papers/bl/jisc-npo67/digital-preservation.html. 7 October 1999.
[15] National Library of Australia. *PANDORA Project: Preserving and Accessing Networked DOcumentary Resources of Australia* (home page). Online. Available: http://www.nla.gov.au/pandora/. 7 October 1999.
[16] University of New South Wales. *Australian Digital Theses Project* (home page). Online. Available: http://www.library.unsw.edu.au/thesis/thesis.html. 7 October 1999.
[17] Consultative Committee for Space Data Systems (CCSDS), CCSDS 650.0-R-1, May 1999. *Reference Model for an Open Archival Information System (OAIS) Draft Recommendation for Space Data System Standards.* Online. Available: http://www.ccsds.org/RP9905/RP9905.html. 7 October 1999.
[18] Consortium of University Research Libraries. *The Cedars Project: CURL Exemplars in Digital Archives* (home page). Online. Available: http://www.leeds.ac.uk/cedars/. 7 October 1999.
[19] CEDARS, Document Number: MGA04, July 1998. *Cedars Project Summary*. Online. Available: http://www.leeds.ac.uk/cedars/documents/MGA04.htm. 7 October 1999.

[20] Russell, Kelly and Sergeant, Derek. "The Cedars Project: Implementing a Model for Distributed Digital Archives." *RLG DigiNews*. Volume 3. Number 3 (1999). Online. Available: http://www.rlg.org/preserv/diginews/diginews3-3.html#feature. 7 October 1999.
[21] van der Werf-Davelaar, Titia. "Long-term Preservation of Electronic Publications: The NEDLIB project." *D-Lib Magazine.* Volume 5 Number 9 (1999). Online. Available: http://www.dlib.org/dlib/september99/vanderwerf/09vanderwerf.html. 8 October 1999.
[22] The British Library. *The British Library Digital Library Programme – Digital Library System Briefing Document.* Online. Available: http://portico.bl.uk/services/ric/diglib/digilib.html. 8 October 1999.
[23] The National Library of Australia. *Digital Services Project.* Online. Available: http://www.nla.gov.au/dsp/. 8 October 1999.
[24] The Royal Library National Library of Sweden. *Kulturarw*^{3} (home page English). Online. Available: http://kulturarw3.kb.se/html/kulturarw3.eng.html. 8 October 1999.
[25] *Eva - the acquisition and archiving of electronic network publications* (home page English). Online. Available: http://renki.lib.helsinki.fi/eva/english.html. 8 October 1999.
[26] Lounamaa, Kirsti, and Salonharju, Inkeri. "EVA - The Acquisition and Archiving of Electronic Network Publications in Finland." *Tietolinja News*. Volume 1 (1999). Online. Available: http://hul.helsinki.fi/tietolinja/0199/evaart.html. 8 October 1999.
[27] Words That Matter Inc. *Electronic Publications Pilot Project (EPPP) Summary of the Final Report*; National Library of Canada, 7 May 1996. Online. Available: http://www.nlc-bnc.ca/e-coll-e/ereport.htm. 8 October 1999.
[28] Electronic Collections Coordinating Group National Library of Canada. *Networked Electronic Publications Policy and Guidelines.* National Library of Canada, October 1998. Online. Available: http://www.nlc-bnc.ca/pubs/irm/eneppg.htm. 8 October 1999.
[29] National Library of Canada. *Electronic Collection.* Online. Available: http://collection.nlc-bnc.ca/e-coll-e/index-e.htm. 8 October 1999.
[30] Personal communications, July and August 1999
[31] Internet Engineering Task Force. *Uniform Resource Names (urn)* (home page). Online. Available: http://www.ietf.org/html.charters/urn-charter.html. 8 October 1999.
[32] The Corporation for National Research Initiatives. *The Handle System* (home page). Online. Available: http://www.handle.net/index.html. 8 October 1999.
[33] International DOI Foundation. *doi The Digital Object Identifier System* (home page). Online. Available: http://www.doi.org/. 8 October 1999.
[34] Online Computer Library Center Inc. *PURL* (home page). Online. Available: http://purl.oclc.org/. 8 October 1999.
[35] Day, Michael. *Issues and Approaches to Preservation Metadata.* Joint RLG and NPO Preservation Conference Guidelines for Digital Imaging, September 1998. Online. Available: http://www.rlg.org/preserv/joint/day.html. 8 October 1999.
[36] eg RLG Working Group on the Preservation Uses of Metadata. *Final Report.* May 1998. Online. Available: http://www.rlg.org./preserv/presmeta.html. 8 October 1999.
Carl Fleischhauer. *Library of Congress-CNRI Experiment Project Proposed Metadata Set.* 12 March 1999. Online. Available: http://lcweb2.loc.gov/ammem/award/docs/nisometa/NISOintr.html. 8 October 1999.
The Making of America II Testbed Project White Paper. Version 2.0 (September 15, 1998). Online. Available: http://sunsite.berkeley.edu/MOA2/wp-v2.html. 8 October 1999.
[37] Stone, Andy, and Day, Michael. *Cedars Preservation Metadata Elements Cedars Project Document AIW02.* CEDARS, 25 February 1999. Online. Available: http://users.ox.ac.uk/~cedars/Papers/AIW02.html. 8 October 1999.
Day, Michael. *Metadata for Preservation Cedars Project Document AIW01.* CEDARS, 3 August 1998. Online. Available: http://www.ukoln.ac.uk/metadata/cedars/AIW01.html. 8 October 1999.
[38] National Library of Australia. *What we need to know – metadata for preserving digital collections.* 1999. Online. Available: http://www.nla.gov.au/preserve/pmeta.html. 8 October 1999.
National Library of Australia. *Digital Collection Management System Logical Data Model.* 23 August 1999. Online. Available: http://www.nla.gov.au/dsp/rft/index.html. 8 October 1999.
[39] Lynch, Clifford. "Canonicalization: a Fundamental Tool to Facilitate Preservation and Management of Digital Information." *D-Lib Magazine.* Volume 5 Number 9 (1999). Online. Available: http://www.dlib.org/dlib/september99/09lynch.html. 8 October 1999.

[40] Fresko, Marc, and Tombs, Kenneth. *Digital Preservation Guidelines: The state of the art in libraries, museums and archives*; European Commission, DG XIII/E, 1998. Online. Available: http://www.echo.lu/digicult/en/study.html.
[41] WGBH. *UPF Home* (home page). Online. Available: http://info.wgbh.org/upf/. 8 October 1999.
[42] Kranch, Douglas A. "Beyond Migration: Preserving Electronic Documents with Digital Tablets." *Information Technology and Libraries*. Volume 17 September (1998).
[43] Heminger, Alan R., and Robertson, Steven B. "Digital Rosetta Stone: A Conceptual Model for Maintaining Long-Term Access to Digital Documents." *Sixth DELOS Workshop Preservation of Digital Information ERCIM Workshop Proceedings - No. 98-W003*. (1998). Online. Available: http://www.ercim.org/publication/ws-proceedings/DELOS6/rosetta.pdf. 8 October 1999.
[44] Norsam Technologies. *HD-Rosetta: Archival Preservation Services.* Online. Available: http://www.norsam.com/rosetta.html. 8 October 1999.
[45] Rothenberg, Jeff. *Avoiding technological quicksand: finding a Viable Technical Foundation for Digital Preservation.* A Report to the Council on Library and Information Resources, January 1999. Online. Available: http://www.clir.org/pubs/reports/rothenberg/contents.html. 8 October 1999.
[46] ibid.
[47] Wiseman, Norman, Rusbridge, Chris, and Griffin, Stephen. "The Joint NSF/JISC International Digital Libraries Initiative." *D-Lib Magazine.* Volume 5 Number 6 (1999). Online. Available: http://www.dlib.org/dlib/june99/06wiseman.html. 8 October 1999.
[48] ibid.
[49] van der Werf-Davelaar, Titia. "Long-term Preservation of Electronic Publications The NEDLIB project." *D-Lib Magazine*. Volume 5 Number 9 (1999). Online. Available: http://www.dlib.org/dlib/september99/vanderwerf/09vanderwerf.html. 8 October 1999.
[50] Bearman, David. "Reality and Chimeras in the Preservation of Electronic Records." *D-Lib Magazine.* Volume 5 Number 4 (1999). Online. Available: http://www.dlib.org/dlib/april99/bearman/04bearman.html. 8 October 1999.
[51] Green, Ann, Dionne, JoAnn, and Dennis, Martin. *Preserving the Whole: a Two-Track Approach to Rescuing Social Science Data and Metadata.* The Digital Library Federation, Council on Library and Information Resources, June 1999. Online. Available: http://www.clir.org/pubs/reports/pub83/contents.html. 8 October 1999.
[52] Ross, S. and Gow, A. 1999 *Digital Archaeology: The Recovery of Digital Materials at Risk.* British Library Research and Innovation Report 108. The other reports are: Beagrie, N. and Greenstein, D. 1998 *A strategic policy framework for creating and preserving digital collections*. British Library Research and Innovation Report 107; Bennett, J.C. 1997 *A framework of data types and formats, and issues affecting the long term preservation of digital material*. British Library Research and Innovation Report 50; The Data Archive, University of Essex 1998 *An investigation into the digital preservation needs of universities and research funders: the future of unpublished research materials.* British Library Research and Innovation Report 109; Haynes, D., Streatfield, D., Jowett, T. and Blake, M. 1997 *Responsibility for digital archiving and long term access to digital data*. British Library and Innovation Report 67; Hendley, T. 1998 *Comparison of methods and costs of digital preservation*. British Library Research and Innovation Report 106; Matthews, G., Poulter, A. and Blagg, E. 1997 *Preservation of digital materials policy and strategy issues for the UK: report of a meeting on the CPA/RLG report, December 1996*. British Library and Innovation Report 41.
[53] Feeney, Mary (ed), 1999 *Digital culture: maximising the nation's investment.* The National Preservation Office, London.