Digital Resources into the Future: digital preservation and the Cedars Project

Kelly Russell

Paper presented at the 22nd UKSG Annual Conference and 4th European Serials Conference, Manchester, April 1999

Preservation and continued access are two major concerns of the electronic era. Choice of strategy, emulation or migration will influence costs, but copyright and management issues may create greater problems. The Cedars (CURL exemplars in digital archives) project is addressing these isuses with the aim of developing preservation strategies.

Kelly Russell is Cedars Project Manager, Edward Boyle Library, University of Leeds, Leeds LS2 9JT
E-mail: libklr@leeds.ac.uk

Overview

Over the last decade academic library services have had to face a number of challenges concerning the introduction and widespread adoption of networked technologies. New communication and information technologies have meant that service delivery, as well as the collections themselves, have sometimes radically altered. The widespread uptake of information technology is not limited to academic libraries but as the main sources of research and scholarship in the UK, academic libraries are increasingly concerned about the implications of providing long-term access to these new scholarly tools. As technology changes and becomes rapidly obsolete, libraries face an urgent situation in attempting to keep these often fragile digital resources 'alive'. This paper will outline some of the most pressing issues facing libraries for preserving or archiving digital materials, as well as look briefly at the Cedars project, funded as part of the eLib Programme to allow the Consortium of University Research Libraries the opportunity to explore this complex area.

Background

The 1990s marked the introduction on a large scale of the concept of the virtual, electronic or digital library. This idea, once a shadowy technical oasis, has moved into the fore and now provides the basis for an environment where access to information (whether digital or not) is greatly enhanced by digital and networked technologies. The Electronic library is not just limited to digital resources: libraries provide digital discovery tools for non-digital collections as well as alternate access to content through digitisation. Resources that are 'within' the collection are no longer limited by physical location. It is an exciting and challenging time.

In the United Kingdom, libraries have been fortunate to receive funding from various sources to kick-start work on the digital

library. The eLib programme is probably the most notable for academic libraries but other sectors like public and school libraries are also reaping the benefits of widespread public interest and adoption of information technology through the New Opportunities Fund and the National Grid for Learning.

However, the rate with which our reliance on digital resources grows is second only to the pace with which those resources change and mutate. As users of technology, we have come to expect upgrades and new versions and to make allowances for these advances. These changes bring improvements, leaving the once cutting edge options to gather dust in the library basement. The speed with which, particularly computer hardware, becomes obsolete is truly breathtaking. Our ability to create (and our tendency to rely on) digital resources far outweighs our current technical and organisational capacity to preserve and provide long-term access to them. In a seminal article in Scientific American, Jeff Rothenburg[1] of Rand Corporation in the US suggests "Digital Information lasts forever – or five years, which ever comes first." Although apparently meant to amuse, this remark is dangerously close to the truth; for some digital resources it may even be an over-estimate.

What is digital preservation?

Libraries have been involved in preservation for as long as they have existed. Although 'preservation' as a discipline or distinct area of work within a library is a relatively recent development, libraries (along with museums and archives) play a key role in preserving a nation's cultural heritage. Although the preservation and conservation of valuable, rare or fragile print material is challenging, the preservation of digital materials will provide the most complex and difficult challenges that libraries have yet tackled. Digital preservation is a complex issue. Although some tasks, once tackled, are revealed in their simplicity, digital preservation on contemplation and exploration becomes more complex. It is worth bearing in mind that sometimes when things seem difficult – they are.

So what do we mean by digital preservation? There continues to be a certain level of confusion about what is meant by this term, contributed to, in part, by the choice of the term 'preservation digitisation' to refer to digital re-formatting/ digital imaging in libraries and archives. Digital preservation can be defined as, "storage, maintenance, and access to a digital object over the long term, usually as a consequence of applying one or more digital preservation strategies". A digital object, in this sense, is defined simply as any resource that can be stored or manipulated by a computer. Such a definition includes both resources that are digitised as well as objects that are 'born digital'. It is important to understand that digital preservation is not concerned only (or even primarily) with digitised resources. Although for the short to medium term academic libraries and archives may continue to focus attention on digital imaging, in the long term, preserving 'digital only' resources presents the most complex challenges because for this material there may be no alternative but to rely on the digital object in future. 'Long term' in this context means through changing technologies and, for the purposes of this article should include anything from ten to 500 years.

Unlike print resources, digital material requires ongoing maintenance to ensure the material is not only stored but made accessible. When information is stored simply as bits (encoded as 1's and 0's), there is no preservation without retrieval. If the material is not meaningfully retrievable, the resource has not been preserved. This is the most important difference between digital and non-digital material. If a book is kept on the shelf, generally this assumes a level of accessibility (it can be opened and read). However, for digital materials, storage of the digital object does not guarantee access. Just because you retain a CD-ROM does not mean it can be opened and accessed. There is little point in preserving a digital object, if access is not possible. Digital preservation is twofold – the storage of the material and ensuring continuing accessibility. It is the latter which is by far the most critical issue and (we assume) the most costly.

Strategies for digital preservation

In a technical sense, digital preservation involves adopting one (or perhaps more than one) of a

number of identified digital preservation strategies. The chosen strategy will reflect the type of resource to be preserved, perhaps the perceived value of the resource within the local or national collection, local expertise and infrastructure, and ultimately the associated costs. Preservation strategies are widely accepted as falling into three broad categories: technology preservation, technology emulation and migration. The first two focus attention on the technology used to create the resource and either seek to preserve that technology or, in the case of emulation, to use current technology to create hard/software platforms which mimic the original technologies. Migration, unlike the others, focuses on moving digital resources through changing technology to ensure it is accessible using current technology. It is worth distinguishing these activities from so called 'refreshing'. Refreshing addresses the threat of media deterioration (e.g. magnetic tape or C.D-ROM) and requires that data be regularly transferred from one storage medium to a fresher one. Any adopted preservation strategy will require refreshing to new media. Refreshing in itself is not digital preservation.

Technology preservation

Preservation of the technical environment by conserving copies of the software and specific hardware is referred to as 'technology preservation'. For example this might require keeping a copy of Windows 95 as well as a machine configured to run it – like maintaining your old record player to continue listening to an album collection. For some digital objects this may be the best solution, at least in the short-term, because it ensure the material is accessible by preserving the access tools as well as the object itself. However, longer term this is more problematic. For example, issues of space and maintenance of the hardware as well as costs may make this impossible in the longer term. This strategy also limits the portability of the resource since they will be dependent on hardware stored in a specific place.

Technology emulation

There are other options, however, for preserving the digital resource which focus on preserving the technical environment without keeping rooms full of old hardware and software. Emulation allows for the use of current technology to mimic original hard/software in order to provide access to the digital object. Although controversial, I believe emulation potentially offers the best solution for very long term preservation of material.

The extent to which the emulation mimics the original technical environment entirely or emulates only those components necessary to access the data remains an issue for debate. Although emulators currently exist for some of the major operating systems, in a library situation, some (if not most) of the material will be on sometimes obscure proprietary soft/hardware chosen by commercial publishers. While other organisations involved in digital archiving may have some degree of control over the types and formats of material they will accept into the archive, libraries are unlikely to be in this position. Where a funding body can make production of resources in a particular format a requirement of funding, libraries are unlikely to have much sway with most publishers who are motivated by commercial interests and not necessarily the desire for longevity of the resources.

It is also worth emphasising that an emulation strategy does not require that an emulator be stored for each archived resource. If it were necessary to store an emulator with each resource deposited, this would mean that the emulators would also be dependent on a particular technical solution and require their own strategy for preservation; the costs of archiving would increase tenfold.

If emulation does not require software/ hardware at the time of deposit in the archive, when is it required? It is also possible to preserve descriptive data about how the technical environment was created in the first place in order to allow for emulation later on if it is necessary. If we preserve enough information about Windows 3.1 then instead of preserving Windows 3.1 itself (or an emulation of it) we can 'simply' re-engineer it again when we need it. This relies on a robust system for preserving the metadata which describes the technical environment.

An emulation strategy means that nothing is done to the original object (it is left simply as a bitstream) and it is the environment which is re-created. The costs of emulation are as yet unknown and it is expected that the costs of re-

creating complex technical environments could be astronomical. However, unlike the technology preservation model described above, the costs fall further along in the resource's lifecyle. Instead of spending money now and for the foreseeable future, by preserving both software and hardware, emulation loads the costs at the far end. If a resource is needed in future, only then are resources required to emulate the necessary technological environment. The need for emulation is therefore determined by demand for the resource so the costs only arise if/when a resource is needed. It is worth also stressing that despite current interest in emulation, this strategy does not mean the material in the archive immediately accessible. If software engineering is required, then this will mean delays in obtaining the resource. It also requires a leap of faith in terms of the power of future technologies and in the abilities of future software engineers. However, of one thing we can be sure, technology will change and it will offer better, cheaper and faster solutions. Emulation may be best for resources for which the value is unknown and where future use of the material is unlikely.

Migration

Where the two options described above, focus on the environment of the object and preserving the resource through re-creating or preserving the necessary operating environment, another strategy for digital preservation is what has been called 'migration'. A report commissioned by the Research Libraries Group and the Commission for Preservation and Access in the US[2], helpfully distinguishes between migration and what has been termed 'refreshing'. The reports suggest that "migration is a broader and richer concept than 'refreshing' .... migration is a set of organized 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. 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."

It is this last strategy in which many libraries and archives are already involved and many believe that this is the most practical approach, at least for the short and medium term. Rather than focus on the technology, this strategy tends to focus on the intellectual content and making it accessible using current technology. However, for resources where it is more difficult to disentangle format from content this is not an easy option. For example, for resources like Microsoft's Encarta or electronic journals which contain bits of sound and video, the content might be inextricably linked to the format or very specific technical environment. Multiple components may require separate migration activities and this can be very complex. Indeed for some multi-media resources migration may not be possible without significant compromises in functionality. In addition, the costs of migration may, in the long run, exceed those costs necessary for preserving either the technology itself or the detailed technical specification which will allow future emulation.

Beyond the technical issues

It is worth dispelling another myth that may prevail. Digital preservation is not only concerned with technical problems. Digital preservation involves consideration of a whole host of non-technical issues. In many cases, the technical issues may pale in comparison to problems associated with copyright, responsibility/ ownership, digital collection management and costs. Preservation or conservation of print materials can in most cases be easily separated from other collection management activities. Indeed preservation/conservation departments in libraries may have little or no connection with acquisitions or systems departments. It is even more unlikely that there are connections between the preservation unit and the computing service – in many institutions, preservation of digital research data may be managed through the computing service and considerable expertise may reside there. The relative isolation of print preservation/conservation, in relation to the increasing need to integrate digital preservation into a range of existing activities (some outside the library service itself), can be problematic for libraries and is perhaps why libraries have, as yet, hesitated to address fully the preservation of digital resources.

Preservation of print materials can be tackled separately. For digital resources, it is not as easy (in fact, it is perhaps dangerous) to divorce preservation of resources from other collection management activities, such as acquisition and maintenance. The format in which a digital resource is acquired and accessioned will influence the way it can be preserved and used in the future. Likewise, a given preservation strategy can have implications for the types and formats of materials that libraries might choose to acquire. These interdependencies highlight the management and organisational issues involved in accepting responsibility for digital preservation – this is not simply a technical problem. Digital preservation in the broadest sense involves a set of organised tasks intended to ensure the continuing accessibility of digital resources. These tasks may be scattered across the library or even across the institution and co-ordinating the activity will be a challenge.

Collection management

Throughout the ages, a great deal of our most valuable and unique scholarly material has been preserved more or less by accident. A manuscript not deemed to be of contemporary value which sits on a shelf can be re-discovered decades or even centuries later, declared an invaluable resource and preserved for years to come. In a digital environment – nothing is preserved by accident. In order to ensure continued access to digital material some deliberate activity is necessary. If we rely on continuing demand (which will ensure access is maintained), then we are taking a short-sighted view and letting the market determine what scholars in the future will be able to use. Selection of material for preservation is a key component of maintaining a digital archive. There has been some work done already in the UK by the Arts and Humanities Data Service (see http://www.ahds.ac.uk) on selection of digital materials for archiving and in the US by at the SunSITE at the University of Berkeley (see http://sunsite.berkeley.edu/Admin/collection.html).

A great deal of work is needed in digital collection management to build on work already done by organisations like the AHDS and to apply it to a library collection context. This extends beyond basic selection of materials to include strategies for selecting dynamic data for archiving, e.g. how often should a database like Medline be archived? Should it be cumulative or should each snap shot be preserved in its own right? Such decisions are collection management decisions and must rely on libraries enhancing, developing or creating policies to deal with these issues. The design and implementation of a digital archive must be led by the policy at the highest level.

Copyright and ownership

In a digital environment ownership of material is not always clear. When a library subscribes to an electronic journal what are they paying for? If they are paying only to access an electronic copy of the journal which is hosted at the publisher's server, then they are potentially left without anyway of guaranteeing long-term access to that material beyond the terms of the licence. Unlike a print subscription, where libraries could (indeed were expected to) bind and keep copies of journals, electronic journals pose new problems. Licences will need to be drafted to take long-term preservation into account.

The technical aspects of digital preservation also have associated rights complications because almost any preservation strategy will involve copying (at the very least) or even some sort of software/hardware re-engineering over time. Those involved in the archiving of electronic material may need to seek special permissions from rights holders to preserve digital resources – particularly those for which there are complex hardware and software dependencies. This is quite apart from negotiating the rights holders over access to the archived material. Even if an agreement suggests that no-one is allowed access to archived material until is it out of copyright, the material will still have to be maintained and retrievable. Permission to access and accessibility are, for digital materials, two separate issues both involving rights negotiations.

Software and hardware functionality offer yet another layer of rights issues. Although the 'intellectual content' of a resource can be considered conceptually separate from the means by which it is accessed, in practice the intellectual content of the resources is frequently inextricably linked to the functionality provided by the

hardware/software platform/configuration. The systems which manipulate the data in order to produce what the end user sees are generally copyright to another organisation. The publisher may have a licence to use the system but is probably not in a position to allow an archive access to the system for archival purposes. Archives and publishers are, therefore, both frequently dependent on third parties, if functionality is not to be compromised.

In many countries world-wide published materials are protected by legislation which requires publishers to deposit copies of published materials in special designated repositories. This may be the national or state library or specific research or academic libraries. These libraries are then given the task (and often associated funding) to preserve these materials into the future. The extent to which this legislation applies in any given country to non-print materials varies. Often film, video or broadcasting material have not been included for reasons of scope and nature of the material itself. There are currently countries where digital materials are included but these remain in the minority. The challenges of preserving digital material remain an obstacle for clear and concise legislation. In the UK and several other European countries there are voluntary schemes for deposit of digital resources but, as yet, very few laws. Before there are meaningful laws however, there must be more research and development to better understand the problems that we will face. It is also important to note that the presence of legislation for legal deposit will not solve the problem for libraries and archives. Legislation of this type applies only to 'published' material. For libraries creating or acquiring digital materials outside this narrow confine, legal deposit legislation is not the panacea. Libraries will still need to understand the resources within their collections for which they take primary preservation responsibilities as well as their role within the institution as a whole in terms of preserving the institutional record in digital form.

In the UK, as a result of a recommendation by a British Library working party the Secretary of State has asked that a code of practice for voluntary deposit of digital materials be drawn up in advance of the eventual legislation which would make deposit mandatory. The British Library is currently working in conjunction with several commercial publishers on this voluntary code of practice. *[Now agreed. Ed.]*

Most electronic resources to which libraries subscribe, or have access, include permission to make back-up copies of the resource, just in case there is damage to the original. Digital preservation, however, is about more than making a backup copy. When material is stored as bytestreams, it can easily be copied from one medium to another for the purposes of back up.

The Cedars Project

At present, libraries lack practical experience in employing these strategies. For this reason the Consortium of University Research Libraries (CURL) has been granted funding to take lead in this area. Funded through JISC's eLib Programme for three years, the CURL exemplars in digital archives (Cedars) project will explore the strategic, methodological and practical implications for preserving and providing long-term access to digital resources. Cedars is led on behalf of CURL by the Universities of Oxford, Cambridge and Leeds and the UK Office of Library Networking (UKOLN) and the main aim of the project is to address strategic, methodological and practical issues and provide guidance in best practice for digital preservation.

It will do this by work on two levels. Firstly, through practical demonstrator projects which will provide concrete practical experience in preserving digital resources. Then, through strategic working groups based on broad concepts or concerns, which will articulate preferences and make recommendations of benefit to the wider community. The main deliverables of the project will be recommendations and guidelines, as well as practical, robust and scaleable models for establishing distributed digital archives.

The project objectives are:

- to promote awareness;
- to identify, document and disseminate strategic frameworks for the development of appropriate digital collection management policies;
- to investigate, document and promote methods appropriate for long-term preservation.