1. INTRODUCTION TO THE ISSUES

1.1 Digital preservation — a contradiction in terms?

Digital information forms an increasingly large part of our cultural and intellectual heritage and offers significant benefits to users. The use of computers is changing forever the way information is created, managed and accessed. The ability to generate, amend and copy information in digital form, to search texts and databases, and to transmit information rapidly over networks has led to a dramatic growth in the application of digital technologies.

At the same time, the experience of addressing the 'millennium bug' in existing software systems, or of losing data through poor management, is beginning to raise awareness of the fragility of this medium, compared to traditional media such as paper. Electronic information is fragile and evanescent. It needs a proactive and strategic approach from its inception to secure its preservation over the longer term.

Although experience in creating and managing specific forms of digital data has been built up over a number of decades in the sciences and social sciences, in many areas it is a relatively new medium where much of the future life-cycle, activities and cost models are currently unknown. These factors have led to increasing concern about the potential loss of our 'collective memory' in the Digital Age.

The situation is encapsulated in a draft policy statement by the Australian National Preservation Office (1997):

> 'As the twentieth century draws to a close, an ever-increasing quantity of information is created, stored, disseminated and networked in digital form. Digital objects, many of which are dynamic in nature, are created by a variety of creators for a number of purposes. Digital objects include data stored in digital form and accessed using electronic equipment. Examples are databases, images, sound, video, documents, etc....
>
> The organisations charged with the responsibility of preserving and making available [our] cultural and intellectual heritage will need to develop a range of strategies to address the preservation of and access to various categories of digital objects. Custodial and non-custodial arrangements will need to be considered both from a preservation and an access perspective and will need to be considered prior to creation if possible and throughout the life of the object.'

The digital world is one of sustained change and flux. Technology is constantly changing; the legal environment is subject to revision and change; digital objects are themselves dynamic. In this context 'preservation' and 'digital materials' seem almost mutually exclusive — how can you preserve something that is constantly changing? And yet digitisation is itself a method of preservation, a way of providing alternative access to the original object. It can be used to preserve and make accessible information not originally produced in this way, especially in the case of originals at risk, such as those printed on acid paper.

The short term benefits of digital objects — manipulation, distribution, duplication, linking — are immense, but the long-term viability of such objects is fraught with difficulty because of the ever-changing technology needed for their storage and use. This apparent contradiction has prompted the search for strategies and processes that will help make digital preservation a reality.

This introductory chapter poses a number of basic questions in order to help provide an overall context for the discussion of preservation issues in connection with digital material. Later chapters examine the issues in more detail and recommend further actions and research.

1.2 Why preserve digital information?

Digital information can be generated by a number of different processes and for different reasons. The information may, for example, exist in a definitive version and be generated by a project or business function with a finite timespan; or it may be dynamic, constantly evolving, generated by a project or business function with no finite timescale. The purpose for which it is created and the reasons why it is preserved may also vary. Non-digital collections may be digitised to improve access or to preserve the information they contain; or collections may be made of existing digital information for future re-use and research.

Whatever the context, preservation is a response to the threat of destruction and loss. Recognising the threat elicits a response, the scale of which is usually in proportion to the value that is placed on the object under threat. Such actions incur costs, which will continue while the threat appears to remain. Funds will be drawn upon, and resources mobilised. Each part of the process will draw different stakeholders into the preservation activity.

With printed, paper-based material there is usually time to consider the best course of action and to make appropriate decisions. The threats to digital information are varied and subtle, and have much shorter timescales than for information on paper. This means that the costs of digital preservation come much earlier and more often, and that substantial costs may be incurred before the value of the object can be realised. Decisions are required, supported by expenditure, to enable resources to be deployed quickly in order to counter the threat of irreversible loss. The resources may involve substantial capital investment as well as specialist labour, both available in the near term only at a premium.

The contest for limited resources and the balancing of conflicting priorities introduce the question of selection — *which material should be preserved?* The solution is not straightforward for any collection developer, though with digital material the threat of loss is ever present and the volume of material requiring attention is growing year by year.

From this other questions grow, concerning the long term viability of any stored information, and the costs and benefits of long-term accessibility.

- What is the rationale for preservation?
- When an object is retrieved from the archive fifty or more years hence, will it still be valuable? Will it still be recognisable, comprehensible and usable?
- Different organisations, for example legal deposit libraries and other research libraries, have their own requirements when retaining material over long periods of time. In each case, what costs are involved? How do these costs apply to an item's life-cycle? How can they share the responsibility?
- Are the benefits measurable? How can they be achieved, and who is responsible for monitoring them?

1.3 Why is digital archiving different from preserving a book?

Books are examples of technological artefacts holding information whose design has matured slowly over hundreds of years. In general, no assisting technology is needed to access the information in a book. Preserving the artefact (simply storing it in good conditions) preserves the information (excepting some cases such as loss of the actual language involved).

Digital technology, by contrast, has only been with us for tens rather than hundreds of years. The technological basis for computing is changing rapidly, advancing far faster than any previous technological developments. Information encoded in digital form — as information objects — is entirely dependent on technology to allow access. There are many ways in which continued access can be threatened: through damage to the medium on which it rests; through loss of the information which describes *how* to access it; or through loss of the computing environment — hardware and software — on which access depends.

Unlike the situation that applies to books, digital archiving requires relatively frequent investments to overcome rapid obsolescence introduced by galloping technological change.

1.4 What should be archived?

It is obvious to most that the digital equivalents of paper publications — books and journals especially — should be treated with the same respect and accorded the same preservation priorities as the paper versions. It is the information content we value, more than the medium or the format, and we must take adequate steps to preserve the newly emerging digital forms of our cultural heritage.

However, our responsibilities do not stop at digital versions of the books and journals currently preserved in their paper form via legal deposit and other mechanisms. There are many other kinds of information that are not currently covered by legislation or by Public Record Office guidelines. Examples include medical records that need to be preserved over time to study disease demographics, and data generated in the course of research. How should these be preserved? The materials preserved must include contextual information by which they can be understood and used correctly. This is

vital, for example, where research data are preserved and used for secondary analysis or to replicate experiments. The Data Archive at the University of Essex is the longest established digital archive in the UK and has long experience of preserving and disseminating such research data.

Research projects in particular produce a huge amount of ‘electronic paperwork’, for example administrative files, electronic research diaries, automatic output from laboratory research instruments, electronic mail. Should these be preserved?

As these few examples illustrate, the volume of digital objects is such that universal preservation would be impossible — some kind of selection is inevitable. However, there are no common selection procedures or agreed guidelines on criteria for selection.

Selection requires qualitative assessment of the value of information. How should this be carried out? The creator of the information can play a key role in assessing its true value, and the research community certainly feels that it should be involved in any selection process regarding research data, but any decision is inevitably subjective. There are risks of censorship, or misjudgement on the part of individuals or organisations. At the same time it is often difficult, if not impossible, to predict the future usefulness of material. Would random selection be a better approach, at least for some types of digital information? One suggestion is for a sampling exercise — a national audit of digital materials, carried out at times specified by a national coordinating body to select those that should be archived according to agreed selection criteria which are evaluated and monitored continuously by the coordinating body.

There is certainly a high management cost associated with a selective policy. However, it is likely that the heavy operational costs of digital preservation techniques will make selection essential.

Different criteria are needed for archiving dynamic products, such as bibliographic databases or discussion forums. Snapshots are one suggestion here. These are versions, linked to date/time stamps. Some datasets require constant updating to maintain their value. Who has responsibility for ensuring the integrity of the updated material? For some non-cumulative databases where individual records are changed or deleted, it may be necessary to keep an audit trail of the changes made to a database over a period of time in order to preserve a comprehensive view of the data held.

1.5 How?

The need to manage the preservation of digital material both immediately and in the long term has encouraged the promotion of a wide range of approaches.

Should material be kept in a standard format or in its original format? Which is more important — functionality (what you can do with the information) or appearance (what it looked like)?

One approach — technology preservation — is to preserve both the original data and the platform necessary to interpret it. However, preserving digital objects in their original format and medium does not guarantee future usability, unless the particular technology on which they depend is also preserved and can be made to work over time.

Migration is the movement of digital objects from one technology to another. It is likely to be a continuous process, with material transferred to a series of new formats over time, and thus potentially an expensive option. In most cases migration is less of a problem from a hardware point of view than that of the software platform.

The technological options are described in more detail in chapter 4.

Cost management principles would suggest that digital material should preferably be held in archives in a standard format, on standard media, and managed by one of a few standard operating systems. Material that does not conform would either have to be processed prior to entering the preservation store or be managed under a different regime with a premium scale of charges.

However, prescriptive standards in the electronic information world have so far failed to achieve full recognition. The emphasis is now on ‘permissive standards’, such as Standard Generalised Markup Language (SGML), which do not tell document creators how to format the document or even what software should be used, but result in an environment that allows exchange of information.

Rather than a prescriptive approach, many of those involved in the creation and management of digital information would prefer to see the development of guidelines and guidance both for specific audiences and for specific types of material.

Certain best practices appropriate for digital preservation can be automated for data creators through the application software they use. This is particularly true with regard to data documentation and metadata, key elements of which can be generated automatically by application software as and when it is used. Accordingly, the development of appropriate software and tools may play a key role in digital preservation.

Prospects for, and the costs involved in, preserving digital resources over the longer term rest heavily upon decisions taken about those resources at different stages of their life-cycle. Decisions taken in the design and creation of a digital resource, and those taken when a digital resource is accessioned into a collection, are particularly influential.

1.6 Whose responsibility?

There are many groups and organisations with some degree of involvement in digital information; they are referred to in this report as *stakeholders*. They bring different perspectives to the need for digital archiving and digital preservation, and their interests are also different. While some are concerned purely with preservation, others

are more interested in access to and re-use of material. For example, re-analysis of data is a central principle of scientific scholarship. Other groups are principally interested in commercial exploitation in the future.

Who are the stakeholders? Some of the groups that have been identified are: authors; libraries; publishers; archive centres; distributors; IT suppliers; legal depositories; consortia; and networked information service providers. We should also include industry and business amongst the stakeholders. For example, pharmaceutical companies need to keep records indefinitely and might be persuaded to contribute to research into preservation methods.

While many stakeholders are involved with data resources at different stages, few have influence over (or even interest in) those resources throughout their entire life-cycle. Organisations with a remit for long-term preservation, for example, acquire digital resources to preserve them and encourage their re-use but often have little direct influence over how they are created. This means that decisions which affect the prospects for and the costs involved in data preservation are distributed across a number of different (and often differently interested) groups.

Where should responsibility for digital archiving lie? At present there are few incentives to preserve data and few requirements to do so.

Responsibility for archiving falls naturally on the creator or owner of the information, who should understand how it works and what its value is. However, ownership of digital data is often unclear, and in these situations it is very difficult for organisations to develop preservation strategies. Ownership also brings other responsibilities, such as making sure that rights management issues and intellectual property are respected. There is a risk that some agencies may disclaim ownership in order to absolve themselves of responsibilities for preservation and liability. Others may use their ownership to prevent actions aimed at preservation for short term reasons.

If the initial creators or owners of digital information fail to meet their responsibilities, then some other organisation, such as a digital archive repository, would have to intervene; legislation would be required for this.

Data creators who attach little or no value to the long-term preservation of the data resources they create (and these are currently in the majority) are unlikely to adopt standards and practices which will facilitate their preservation, especially where doing so would involve extra costs. We need to make them aware of the benefits of preservation in a manner which appeals to their interests.

Chapter 2 discusses more fully the rights and responsibilities of stakeholders, while chapter 3 suggests a policy framework to assist those involved in the creation and preservation of digital resources to formulate their own data policies.

1.7 Who pays? How much will it cost?

The widespread and growing use of digital objects means that their preservation is a general problem throughout society. Failure to preserve them adequately will damage future scholarship and weaken the cultural heritage.

But who should pay for digital preservation? Is the system for print publications — legal deposit through the British Library and other deposit libraries — suitable for digital archives? It is widely felt that digital archives should be funded by the government, ultimately by the taxpayer. An alternative approach might be for communities with a common interest to decide to fund digital preservation on a shared basis. Funding could be shared among various sources, with publishers, creators of information (including the academic community), libraries and users all making a contribution. Other possibilities include private sector involvement, funding by charitable trusts, as in the USA, and international collaboration through the European Union.

Should users pay fees for access? Or should it be free at the point of use? The possibility for generating income exists but is also unquantified.

Overall, digital archiving is a cost-unknown venture. We need to establish more precisely how much it will all cost, by constructing different working models based on different ways of doing it. Any strategy for long-term preservation must also take into account the possibility that the level of resources that may be devoted to digital archiving will not be available over the long term.

Chapter 5 discusses the costs involved in digital preservation and proposes a model that can be used to compare the costs of different methods of preservation.

1.8 Who has access?

Access involves both technical issues (ensuring that digital objects are maintained in a usable form) and legal issues (establishing ownership and protecting copyright). For an archive, a digital object must be technically accessible both when access is legally permitted and when it is not.

It is also a contentious issue. There seem to be two distinct groups involved, each with different motivations: those whose primary interest is access to digitised materials on as wide a scale and at the lowest cost possible; and those who want preservation with no, or very restricted, access. The challenge is to find a way of satisfying both groups.

What is needed is a series of generic rules on access that can be adapted by negotiation in specific cases. There is also likely to be a number of new operating arrangements, enabled by technology, to secure appropriate access.

A key issue is how to provide users with access to material while protecting the interests of copyright holders. Is ‘fair dealing’ possible with electronic publications? Can copyright apply to transient data? Several solutions have been suggested: providing recompense to the copyright holder for use while material is current; licensing arrangements; or the use of metadata to embed the details of the copyright

holder in the document. IT mechanisms may also be deployed which ensure that unauthorised copying is impossible. For example, an Electronic Copyright Management System (ECMS) may produce a watermark — a mark embedded in a document which will show on all printouts to identify the publisher, even if the text is edited.

There remain problems, though, in establishing ownership of digital material. In many cases it is very difficult to identify exactly who is the rights holder. The enormous variety of agencies publishing and distributing digital objects, along with the numerous and ill-defined roles of the creators of digital objects, make for a confused and complex situation.

Apart from the fraught area of copyright, authors and owners of digital information have moral rights which must also be respected, including integrity (protection against corruption of the work), the right to be named as author, and protection against misattribution. It is not yet clear how moral rights will affect preservation.

Both users and owners of digital material require guarantees of security and authenticity; the assurance that, once archived, an item cannot be changed.

1.9 What kind of strategy?

At present, the short-term focus on cost-efficiency during data creation is dominant. It is up to relevant organisations to take an active role in publicising to other stakeholders the value of the long-term preservation of selected digital resources, and to demonstrate the benefits of any additional investment during data creation in terms of efficiencies and use later in the life-cycle of the resource.

Digital preservation remains relatively undeveloped. The UK lacks a strategy for the long term preservation of digital information on a scale sufficiently large to support future scholarship and research. A strategy for digital preservation is part and parcel of any national information policy and it should be integral to any investment in digital libraries and information superhighways. How should such a strategy be formulated? What form should it take?

The government has now accepted the principle that digital archiving should be subject to legislation in the same way as, or as an extension of, legal deposit for printed material. Without a legal deposit system it would be unclear how any authority would ensure that originators preserve their material, or make partnerships to preserve it. Even with such a system, there will still be many thousands of originators whose work is not ‘published’ and which therefore falls outside the scope of legal deposit legislation. It will be difficult to ascertain whether they are all preserving their material properly. As a fail-safe mechanism, archives could be given rights for ‘aggressive rescue’, where organisations were seen to be failing in their responsibilities, but policing such a system would be difficult and expensive.

Should the system be centralised or distributed? Centralised storage offers the benefit of economies of scale. However, a central body might be seen as a threat to the

independence of existing agencies. Different archiving groups may be formed where there is common interest in preserving a cluster of information, as defined by the stakeholders.

The nature and scale of long-term digital preservation are such that no single agency is likely to be able to undertake the role of preserving all digital materials within its purview or the necessary research and development in this field. Cooperative agreements and consortia will be required. These will need to address a wide range of issues including, for example, the division of responsibility for different subject areas or materials, selection of material, the degree of redundancy which may be desirable for preservation or multiple locations for access, funding, and different national or regional needs.

Preservation is widely regarded as being for the common good, but there are high costs involved. What part will market demand play in deciding what will be archived? Should it be left to market forces alone? Can we harness market forces for the public good by making it economically rewarding to deposit data, to encourage people to comply as they realise the value locked up in their data? The Task Force on Digital Archiving, created in the USA by the Commission on Preservation and Access and the Research Libraries Group (Waters and Garrett, 1996), commented that:

> ‘Without the operation of a formal certification program and a fail-safe mechanism, preservation of the nation’s cultural heritage in digital form will likely be overly dependent on marketplace forces, which may value information for too short a period and without applying broader, public interest criteria.’