**British Library Research and Innovation Report 107**

# A Strategic Policy Framework for Creating and Preserving Digital Collections

***A Report to the Digital Archiving Working Group***

***by***

***Neil Beagrie and Daniel Greenstein, Arts and Humanities Data Service Executive***

***King's College, London***

British Library Research and Innovation Centre

1998

This study is part of a programme funded by JISC as a result of a workshop on the Long Term Preservation of Electronic Materials held at Warwick in November 1995.

The programme of studies is guided by the Digital Archiving Working Group, which reports to the Management Committee of the National Preservation Office.

The programme is administered by the British Library Research and Innovation Centre.



RIC/G/412

ISBN 0 7123 9714 0

ISSN 1366-8218

British Library Research and Innovation Reports may be purchased as a photocopy or microfiche from the British Thesis Service, British Library Document Supply Centre, Boston Spa, Wetherby, West Yorkshire LS23 7BQ, UK.

"The study presents thirteen recommendations in the areas of long-term digital preservation, standards, the policy framework, and future research. Six case studies highlight some of the real-life considerations concerning digital preservation. At a time when content providers and libraries are racing headlong toward digitization of information resources, this study provides critical guidance." Internet Scout Review, Volume 5, Number 2, 8 May 1998

# A strategic policy framework for creating and preserving digital collections

Version 4.0, 14/7/98
Final Draft
Neil Beagrie and Daniel Greenstein Arts and Humanities Data Service Executive
King's College London
Strand
London WC2R 2LS

**Contents:**



## Preface

This study is part of a programme funded by the Joint Information Systems Committee (JISC) on behalf of the Higher Education sector in the UK, following a workshop on the Long-term Preservation of Electronic Materials held at Warwick in November 1995.
The programme of studies is guided by the Digital Archiving Working Group, composed of members from UK Higher Education Libraries, Data Centres and Services; the British Library; the National Preservation Office; the Research Libraries Group; and the Publishers' Association. The Group reports to the Management Committee of the National Preservation Office.
The programme is administered by the British Library Research and Innovation Centre.
This study has been researched and written by Neil Beagrie (Collections and Standards Development Officer) and Daniel Greenstein (Director) of the Arts and Humanities Data Service (AHDS) Executive. The AHDS is funded by JISC on behalf of the UK Higher Education community to collect, manage, preserve, and promote the re-use of scholarly digital resources. Further information on the AHDS and its constituent Service Providers is available from the AHDS web site http://ahds.ac.uk/.

## 1. Structure and Contents

The report addresses the critical issue of developing a strategic policy framework for the creation and long-term preservation of those digital resources which will form our future cultural and intellectual heritage. It consists of the following sections:

- Executive Summary and Recommendations
- an introduction consisting of two parts - the background to the study, its aims, methodology, and relationship to other initiatives; and secondly an introduction to the issues in creating and preserving digital information, the importance of digital preservation and the policy framework;
- a high-level presentation of the framework identifying how policies need to address the key stages in the life cycle of a digital resource, the inter-relationships and dependencies between each stage, and how these are influenced by the legal and business environment within which the digital resource is created, used and ultimately preserved;

- case studies, demonstrating how issues identified in the framework have been addressed by organisations in the different business environments encountered during the study. The case studies provide a synthesis of information from a number of separate structured interviews, arranged to reflect similar business missions and roles. Each case study identifies common approaches and issues, and provides a detailed examination of each stage in the framework and of the policies and practices adopted by the interviewees;
- a summary of best practice and standards in implementing the framework;
- a bibliography and list of further sources of and references for the study (including World Wide Web references and literature on standards, current research, and ongoing projects which will provide further guidance on specific sectors, media, and issues relevant to the effective implementation of the framework and for supporting digitisation and preservation programmes);
- appendices with the interview questionnaire and draft framework.

## 2. Executive Summary and Recommendations

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 being created, managed and accessed. The ability to generate, easily amend and copy information in digital form; to search texts and databases; and to transmit information rapidly via networks world-wide has lead to a dramatic growth in the application of digital technologies.

At the same time the great advantages of digital information are coupled with the enormous fragility of this medium over time compared to traditional media such as paper. The experience of addressing the Year 2000 issue in existing software systems, or data losses through poor management of digital data are beginning to raise awareness of the issues. Electronic information is fragile and evanescent. It needs careful management from the moment of creation and a pro-active policy and strategic approach to its creation and management to secure its preservation over the longer-term. The cost structure for securing the cultural and intellectual work of the digital age will be notable and has to be built in at the beginning if these costs are to be minimised and that investment effectively applied. There will be many stakeholders and interests in a digital resource over a period of time. A strategic approach is needed to recognise, address, and co-ordinate these interests and secure the future of digital resources.

The framework elaborated by this study provides strategic guidance to stakeholders involved with digital resources at various stages of their life cycle. Although its aim is to facilitate awareness about practices which may enhance the prospects for an d reduce the cost of digital preservation, it is useful for anyone involved in the creation, management, and use of digital resources. Key issues which should be addressed by stakeholders in order to identify and select appropriate and cost-effective practices may be identified for each stage of the digital resource's life cycle and are summarised in the report.

The study suggests that the 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.

The study also suggests that different (and often, differently interested) stakeholders become involved with data resources at different stages. Indeed, few organisations or individuals that become involved with the development and/or management of digital resources have influence over (or even interest in) those resources throughout their entire life cycle. Data creators, for example, have substantial control over how and why digital resources are created. Few as yet extend that interest to how those resources' are managed over the longer term. In some cases they cannot, particularly where resources are not available or allocated for this task. Organisations with a remit for long-term preservation, on the other hand, acquire digital resources to preserve them and encourage their re-use but often have little direct influence over how they are created. One consequence, is that decisions which affect the prospects for and the costs involved in data preservation are distributed across different (and often differently interested) stakeholders. Although stakeholders have a clear understanding of their own involvement with and interest in digital resources, they have less understanding of the involvement and interests of others.

Further, they may have little or no understanding of how their own involvement influences (or is influenced by) them, or awareness of the current challenges in ensuring the long-term preservation of the cultural and intellectual heritage in digital form.

The use of standards throughout the life cycle of the digital resource was emphasised by all respondents. Their application variously ensured that data resources fulfilled at minimum cost the objectives for which they were made. They also facilitated and reduced the cost of data resources' interchange across platforms and between individuals. Standards' selection and use, however, was highly contingent upon where in its life course any individual or organisation encountered a digital resource, and on t he role that that individual or organisation played in the creation, management, or distribution and use of that resource.

The study finally suggests that funding and other agencies investing in the creation of digital resources or exercising strategic influence over the financial, business, and legal environments in which they are created can be key stakeholders. Where they recognise the long-term value of resources created under their influence, their perspective facilitates an interested overview of how those data resources are handled through the different stages of their life cycle. At the same time, their strategic influence may enable them to dictate how those resources are handled. In the case of the Natural Environment Research Councils (NERC), that perspective and influence have been brought to bear effectively with regard to the preservation of NERC-funded data resources. Organisations which retain digital information to document their activities and for other purposes, may have the same perspective and the same degree of control as is evident in the policies and guide-lines available from the UK's Public Record Office and the National Archives and Records Administration of the United States.

A number of observations and recommendations arise from these findings:

**1. Long-term digital preservation**

1.1. Digital preservation is an essentially distributed process including a range of different (and often differently interested) stakeholders who become involved with digital resources at particular phases of their life cycle. To increase the prospects for digital preservation and reduce their costs, different groups of stakeholders need to become more aware of how their particular involvement with a digital resource ramifies across its life cycle.

1.2. Data creators who attach little or no value to the long-term preservation of the data resources they create are unlikely to adopt standards and practices, which will facilitate their preservation. This is particularly true where those standards and practices are different from or more costly to implement than those which promise the cost effective development of a data resource capable of fulfilling its intended use. Accordingly, the awareness-raising suggested above needs to be addressed toward data creators in a manner which appeals to their interests.

1.3. Use of the strategic framework and guidance proposed in this study will assist stakeholders in identifying issues and dependencies and could assist in raising awareness of the strategic issues across the range of stakeholders we have identified.

1.4 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 gene rated 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.

1.5 Several stakeholders are involved in managing data over the longer term, including data banks, institutional archives, and academic data archives. Further research and development initiatives are apparent in the library and cultural heritage sector s, though particularly in the former. Despite their different aims, and the different business, funding, and legal environments in which they work, these stakeholders share a great deal in common. None the less, there were few channels established to facilitate their inter-communication. Cross-fertilisation and information sharing is crucial to these stakeholders, some of whom have 30 years and more of

highly relevant data management experience. Particular attention should be paid to the experience of the data banks and the institutional archives - experience which is often overlooked in other current research and development activities.

1.6 A number of the organisations interviewed for the study have begun to implement pro-active strategies to influence the life cycle of digital resources and manage the process. We have used the term "remote management" to describe the processes observed to manage "active" or "dynamic" resources, or to contract for specialist skills and facilities. Remote management appears to be an widespread response to a distributed process and best practice in its use should be developed and encouraged.

1.7 Funding and other agencies which invest in the creation of digital resources creation or have a strategic influence over the financial, business, and legal environments in which that work takes are best positioned to facilitate consideration of long-term preservation over the life cycle of the resource.

1.8 The nature and scale of long-term digital preservation will encourage co-operative activity between organisations. 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, and co-operative agreements and consortia will be required. These agreements and consortia will need to address a wide-range of issues including for example, the division of responsibility for different subject areas or materials, the degree of redundancy which may be desirable for preservation or multiple locations for access, funding, and different national or regional needs.

**2. Standards**

2.1 Information about standards are currently documented by organisations which identify, document, and promote them, as is evident from the list of relevant standards agencies supplied in the bibliography. Less information is available about how a constellation of standards and methods may be applied effectively to a digital resource at various stages of its life cycle in order to achieve very specific and clearly articulated aims. It is a recommendation of this study that such "best practices" be identified and, where necessary, documented, and that integrated access to them be provided in a meaningful way.

**3. The Policy Framework**

3.1 To implement the framework, stakeholders are recommended to assess the issues pertaining to them, but also to understand how their approach to those issues may have ramifications for the data resources which come under their remit and for other stakeholders which have been or may become involved with them at other stages of their life cycle.

**4. Further Work**

The following further work is recommended to elaborate issues addressed in this study:

4.1. Further research is required into the data policies and practices as implemented by some stakeholders. In particular, research is recommended into the policies and practices of business archives and electronic publishers.

4.2. The study uncovered interest in emulation and technology preservation as a preservation strategy for some digital resources but little evidence of any detailed research into the cost and conduct of those strategies. In the United States, research in this area is currently being conducted by Jeff Rothenberg. Such research is recommended as a matter of priority.

4.3. The study uncovered stakeholders with long-standing experience of different data creation and management policies and practices. The cost models associated with these different policies and practices could have been constructed only they were outside the scope of the current study. Such cost models should be constructed as a matter of priority.

4.4 Several interviewees stressed the importance of demonstrating the cost-effectiveness of a higher initial investment in standards and documentation at the data creation phase to meet the requirements long-term preservation, and thus allowing use of the resource over a longer period. This concept was seen to be required to address what they perceived as a dominant short-term focus on cost-efficiency during data creation. We recommend that relevant organisations actively publicise the value of the long-term preservation of selected digital resources to other stakeholders, and demonstrate the benefits of any additional investment towards long-term preservation during data creation in terms of efficiencies and use later in the life cycle of the resource.

# 3. Introduction

## 3.1 Background

*The Programme of Preservation Studies*

In 1995 a workshop was held at Warwick University to consider The Long-Term Preservation of Electronic Materials (Fresko 1996). The workshop was convened to consider issues raised in the draft report of the Task Force on archiving of digital information commissioned by the Commission on Preservation and Access and the Research Libraries Group in the US and published in the following year (Garrett and Waters 1996). The workshop made a number of recommendations for further investigation and research within the UK and the Joint Information Systems Committee subsequently agreed to fund a research programme, developed in conjunction with the National Preservation Office and administered by the British Library Research and Innovation Centre.

*Aims of this Study*

This study aims to provide a strategic policy framework for the creation and preservation of digital resources, and to develop guidance based on case-studies, further literature and ongoing projects which will facilitate effective implementation of the policy framework. The framework itself is based upon the stages in the life cycle of digital resources from their creation, management and preservation, to use, and the dependencies and inter-relationships between these stages and the legal, business and technical environments in which they exist. The case studies and other guidance incorporated in the report have been developed to illustrate how the framework can be used and applied by different agencies who may have different roles and functions, and in some cases direct interests in only part of the life cycle of the resource.
The intended audience for the study therefore encompasses all individuals and organisations who have a role in the creation and preservation of digital resources from the funding agencies, researchers and digitisers and publishers, through to the organisations which may assume responsibility for their long-term preservation and use.

Through this framework and guidance the study specifically aims to:

- provide guidance in formulating policies which are appropriate for the purposes of data creation, management, and long-term preservation;
- assist agencies in designing digitisation programmes which maximise their cost effectiveness and fitness for purpose over the life cycle of the resource;
- inform strategic planning amongst agencies which invest in the creation and/or collection of digital information resources and seek in some way to ensure the long-term viability of those resources;
- help raise awareness of the strategic issues, dependencies, and need for co-operation between the different stakeholders and agencies identified in the study;
- select and bring together case studies and literature on standards, current research, and ongoing projects which will provide further guidance on specific sectors, media, and issues relevant to the effective implementation of the policy framework and of supporting digitisation and preservation programmes;
- provide a launch pad for more detailed investigations into any of the issue areas

which the framework addresses.

*Methodology*

The study was carried out by Mr Neil Beagrie (Collections and Standards Officer, AHDS Executive) and Dr Daniel Greenstein (Director, AHDS Executive) between December 1997and March 1998. It was based upon traditional desk-based research methods and on fifteen structured interviews. The former involved extensive and growing literature, much of it available freely on the World Wide Web, and also in subscription-based print and electronic journals, and trade association newsheets. Crucially it also too k account of the policies and programmes which large-scale digital preservation and digital collection development initiatives are beginning to provide in some "published" format.

In preparation for the study interviews, a questionnaire and draft framework document [see Appendix 1]; the proposal for the study; and the AHDS webpage pointing to preservation resources and projects, were mounted on the AHDS website. Interviewees were sent details of these documents and requested to consider them in advance of the interview.

Structured interviews, conducted in person or over the phone or by email, involved senior data managers and specialists working in organisations both in the UK and overseas with experience in digitisation, data management or the long-term preservation of digital information resources. Interviewees were selected to provide a wide cross-section of experience of different media types, and experience in different sectors such as national museums, archives, and libraries; university computer centres and data archives; scientific data centres; and research libraries.

We are indebted to the members of the Digital Archiving Working Group, those who commented on the consultation draft of the study report, and to the following individuals and organisations who participated in the interviews and contributed extensively to the study:

- Adrian Cooper and Alan Seal, Victoria and Albert Museum
- Alice Grant and Sue Gordon, National Museum of Science and Industry
- David Giaretta, Rutherford Appleton Laboratory and ISO CCSD Panel 2
- George Darwall, Natural Environment Research Council
- Mirjam Foot and Mike Alexander, The British Library
- Ian MacFarlane and Susan Healy, Public Record Office
- Peter Graham, Rutgers University New Jersey
- Alex Reid, University of Oxford Computing Service
- Kevin Ashley, University of London Computing Centre
- Simon Harden, British Film Institute
- Sandy Buchanan, Scottish Cultural Resources Access Network (SCRAN)
- Jasmine Cameron, Jan Fullerton, Margaret Phillips, Debbie Campbell, National Library of Australia
- John Price Wilkin, University of Michigan
- Margaret Adams, Center for Electronic Records, National Archive and Record Administration of the United States
- Sheila Anderson, Mike King, Peter McKay, Ken Miller, Kathy Sayre, Data Archive, University of Essex

The literature survey and interviews were used to:

- review, amend, and ultimately validate the areas identified in the draft framework;
- identify and document case-studies of the practices adopted within these areas by agencies with significant experience in digitisation, management, or long-term preservation of digital information;
- identify further instructional and methodological literature on standards and current research for specific sectors, media, or issues, relevant to the effective implementation of the policy framework

Information from the literature survey has been incorporated in to the chapter on bibliography, resources and references for the study. Similarly, information from the structured interviews has been incorporated in to the chapter of case studies.

Further review and consultation with professional organisations, specialists and institutions with an interest in its contents was sought by: circulating copies to AHDS Service Providers, other stakeholders, and the study interviewees; and by placing the draft on the AHDS webpages and inviting further input and comments via appropriate email-lists and correspondence.

*Relationship to Other Initiatives*

This study has been undertaken as part of a programme of studies in the UK and should be seen as part of an integrated series of research co-ordinated by the UK's Digital Archiving Working Group. The study will provide a resource for new initiative s within the Higher Education sector such as CEDARS piloting digital preservation in electronic libraries, and for existing initiatives such as the Arts and Humanities Data Service and the Data Archive who are promoting the access and preservation of other digital resources.

At the same time the study has taken a cross-sectoral approach and drawn on the expertise of the library, data, archive, and museum sectors. During the course of the study we have established contact with a wide range of initiatives in these sectors, which we believe to be complementary and desirable to maintain.

For example Panel 2 of the Consultative Committee for Space Data Systems within ISO is developing a draft reference model for an Open Archival Information System (OAIS) for the long term preservation of digital information obtained from observations of the terrestrial and space environments. The reference model aims to provide a framework and common terminology that may be used by Government and Commercial sectors in the request and provision of digital archive services. Although primarily aimed at the space and earth observation communities, the model recognises that it could be extended to other communities. The chair of the UK Working Party for the OAIS standard has been interviewed as part of this study. We believe the work undertaken by the ISO committee on behalf of the Space and Earth Observation communities is complementary to our own and that maintaining dialogue with this initiative would be mutually beneficial.

### 3.2 Significance and Role of the Framework

*Creating and Preserving Digital Information*

Computerisation is changing forever the way information is being created, managed and accessed. The ability to generate, easily amend and copy information in digital form; to search text and databases; and transmit information rapidly via networks world-wide, has led to a dramatic growth in the application of digital technologies to all areas of life. Increasingly the term "Information Age" is being used to describe an era where it has been estimated we have created and stored one hundred times as much information in the period since 1945 as in the whole of human history up to that time.

This new environment poses many opportunities and challenges for those who are involved in creating, preserving or using information in a digital form:

- The content is stored as a series of bits ('1's or '0's) which require hardware and software to retrieve a stream of bits and interpret them as character sets, fields of information and formats, before displaying the information in a visual or audible form which can be understood by the user. Unlike the printed word which can remain accessible over hundreds of years to different generations of users, digital information cannot be understood without the technical data stored with it. This technical data is normally concealed from the user and needs to be preserved and migrated with the content by embedding it in accompanying metadata and documentation.
- With the current rapid changes and evolution in hardware and software, digital information needs active management from its inception if it is to survive and be kept accessible across different technological regimes.
- The magnetic and optical media on which digital information is stored are impermanent and cannot be relied upon for preservation of their contents for more than a few years or decades. In comparison, information on paper or microfilm produced to

appropriate standards and maintained in appropriate environmental conditions can survive for hundreds of years. Digital information therefore needs more active management and intervention to maintain it than other media.

- Digital data has allowed the development of new types of information: dynamic resources which are constantly changed and updated, e.g. databases; interactive resources which are highly contingent on their hardware and software environments for the nature of the experience they create, e.g. games software; or the hyper-linked documents and images found on the Web.
- The provenance and context of digital information is not transparent and easily understood by the user. Unlike traditional paper media the context of a particular digital document is not conveyed intuitively. In traditional record or filing systems a memo or document version will be grouped and positioned in context to other related documents and its provenance and context can be understood by a user. With digital information provenance and context needs to be explicitly captured and documented as it is created to replicate information conveyed by the arrangements and structures used for traditional media .
- The ease with which digital information can be copied and amended is one of its greatest benefits. At the same time however this poses problems for the user in determining whether the document is original and not subsequently altered intentionally or otherwise; or when many versions of a document exist in determining their relationship to each other. The fixity and authenticity of digital information is therefore an issue.
- The legal framework in which digital information is used is often distinct from other media. Increasingly digital information objects are not "owned" by a user or repository but licensed from their creators and their use governed by contractual terms. The rights and terms attached to a digital object when it is created or acquired may fundamentally control how or whether a repository can preserve it or make it accessible to future users.
- The substantial volume and rate of growth of digital information places an increased importance on creating resources which are fit for purpose and cost-effective over their full life cycle. It also emphasises the importance of the ability to select, retrieve and store this information in the most cost-effective and efficient manner possible both to maintain budgets for these activities and prevent systems and users becoming overloaded by information.

*The Importance of Preservation and Access*

Digital information forms an increasingly large part of our cultural and intellectual heritage and offers significant benefits to users. At the same time preservation and access to this information is dependent on impermanent media and technologies ; retaining metadata on the provenance and context; and retaining the authenticity and content of the resource. To assess and retain the content of digital information over time remains a substantial challenge. Converting the digital content to analogue format with known long-term preservation qualities can be a potential solution in some cases and "hybrid" microfilm storage/digital access solutions for some digital information have been explored. Similarly organisations have often used a paper print-out to provide elementary back-up. However with the increasing complexity of electronic information such strategies can be limited and electronic content and functionality can be lost. Increasingly we need to preserve the information in electronic form. 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 and have prompted further research into the long-term preservation of digital information and maintaining future access to it.

Substantial digital preservation initiatives are currently underway in Britain, for example at the British Library, the Public Record Office, the Data Archive, the Natural Environmental Research Council, and the Arts and Humanities Data Service. Further initiatives are contemplated by the Joint Information Systems Committee, by the British Library, and by individual heritage and educational agencies which find themselves increasingly concerned with long-term preservation

of the digital information resources which they are helping to create or archive. Growing British interest in digital preservation is complemented and shared internationally for example by the work of the Commission on Preservation and Access, the Research Libraries Group, and the National Archives and Records Administration in the US; by the National Library and National Archives of Australia; and by various initiatives in Europe such as the DLM-Forum, and elsewhere.

*The Importance of a Policy Framework*

The challenges posed by digital information have increasingly led to recognition of the inter-dependence between the stages of creation, use and preservation of digital resources and the importance of the legal and economic environments in which they operate. The potential volume of information which could be acquired or digitised, and the need to make the most cost-effective use of limited resources, have emphasised the need for selection, standards and co-operation between different organisations . Organisations are developing internal policies for the creation, management, and preservation of digital resources and increasingly are sharing their experience in this field.

A key part of this shared experience has been the recognition of the importance of the life cycle of digital resources and the complex inter-relationships between different practices which may be adopted to create, use or preserve them. Digital preservation is crucial as part of a series of other issues which effect the creation, storage and use of a resource. These issues are all inter-dependent and have suggested the need for an integrated policy framework to develop a cost-effective approach resource creation, preservation and use.

An integrated policy framework may also assist funding agencies in maximising their scholarly and financial investment in the creation of primary and secondary data resources, and data creators in maximising the cost-effectiveness, fitness for purpose, and design, of their digitisation programmes.

This study aims to identify current practice, strategies and literature relating to the creation and preservation of digital information and to provide the integrated policy framework and guidance, which many believe are crucial to long-term preservation of digital resources.

# 4. The Policy Framework

## 4.1 The Development of the Policy Framework

The starting point for this study as outlined in the methodology (see section 3.1) was the draft policy framework. This represents selected elements of a generic collections policy developed for the Arts and Humanities data Service (AHDS), a distributed national service and collection established by the Joint Information Systems Committee of the UK's Higher Education Funding Councils. The development and implications of the AHDS collections policy and study framework has been described elsewhere by the authors (Greenstein 1997, Greenstein 1997, Beagrie forthcoming)

The AHDS is a multi-disciplinary service with five service providers covering archaeology, history, literary and linguistic texts (the Oxford Text Archive), performing arts, and the visual arts, with a remit to collect, catalogue, manage, preserve, and promote the re-use of scholarly digital resources. Its collections policy was therefore developed to cover a wide-range of subject disciplines and different digital media, and provided a valuable starting point for the study.

The AHDS collections policy applies the concept of the life cycle of a digital resource, which has been widely used in the records management and archival professions (e.g. European Commission 1997a, 1997b) as part of the framework used for its construction. The policy framework outlined below also employs the concept of the life cycle of a digital resource. It has extended and enriched the draft framework to reflect the perspectives, experience and roles of other stakeholders who can be involved in the creation and preservation of digital resources, as identified in the study interviews and the literature search.

## 4.2 How to use the Framework

The framework outlines the three main stages (creation, management / preservation, and use) in the life cycle of a digital resource, the role and functions of different generic stakeholders within this, and the inter-relationships between each stage and the implications for preservation of those resources with long-term cultural and intellectual value.
The inherent properties of digital resources mean that the processes of data creation and long-term preservation will involve a wide range of individuals and institutions which have a short-term or even indirect interest, as well as including institutions with a traditional role in these processes(see 4.2 Applicability and Scope below). The framework therefore identifies the roles and functions of different generic stakeholders so that individuals and institutions can see how they and others fit into the framework. Use of the framework may thus facilitate effective collaboration between different stakeholders over the life cycle of the resource. The life cycle of the resource is also heavily influenced by the legal and business environment, so the framework explains the influence of these factors and how they may shape the creation, management, and use of the resource.

To use the framework in drafting strategic policies or implementation guidance the user should "walk through" the framework considering the aims they are trying to achieve, the issues and other players at each stage in the life cycle of the resource, and how they will be influenced by the legal and business environment in which they operate. The framework therefore effectively provides a high-level checklist which individuals and institutions can use to develop policies and guidance which they will tailor to their specific function or role and environment. In so doing they will also identify the implications across each stage, and the impact on or made by other players involved. The overall effect should be to provide policies and implementation strategies where the cost/benefits have been fully explored and strategic partners or dependencies identified.

The Case Studies are intended to illuminate this process further by providing a synthesis of the existing practice, policies and implementation strategies of those interviewed for the study. The Case Studies show how issues have been approached in practice and how different organisational missions shape approaches to creation and preservation of digital resources. This can then be elaborated further by reference to the additional bibliography and references.

## 4.3 Applicability and Scope

The study is concerned with the creation and long-term preservation of our cultural and intellectual heritage in digital form.
For the purposes of digital preservation, long-term can be defined as beginning when the impact of changing technology such as new formats and media needs to be addressed and extending indefinitely thereafter. In a digital environment, the framework an d preservation will therefore include institutions with a traditional interest in long-term preservation but will also extend to a wider range of individuals and institutions which have a short-term or even indirect interest in this process.

The digital information covered by the framework can be the primary form of the data, surrogate versions of primary information held in digital or physical form, or the metadata for collection management of these objects. The framework recognises that digital media are new, distinctive, and require new approaches to their preservation. At the same time it recognises that these approaches may need to be integrated with those for other media and, where relevant, should draw on the existing and extensive professional experience in managing them. It recognises that individuals and organisations may be responsible for hybrid resources consisting of a mixture of digital and other media, or solely focused on information in a digital form. The framework will therefore be applicable to those seeking to extend and modify existing policies for traditional collections to include digital information and for those developing data policies for purely digital collections.

Digital information can be generated by a number of different processes and for different

purposes each of which is considered by the study. The information may 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, and generated by a project or business function with no finite timescale. The purpose for which it is created and preserved may also vary from digitisation of existing information to improve access and/or preservation of existing collections; to the collection of existing digital information and its preservation for future re-use and research.
The chapter of case studies introduces a range of stakeholders and organisational roles in the creation, management and preservation of digital resources encountered during the study. Individual institutions need not be confined to a single role but normally a single role was found to have a greater influence on its approach to data creation, management and preservation, and use. These roles are described in greater detail later in the report and can be summarised as follows:

- funding agencies;
- "digitisers" including research-oriented agencies and individuals, many library and cultural heritage organisations, and publishers
- "data banks" archiving digital information at the bit level usually under contract for a third party;
- institutional archives managing unique electronic records generated by a single organisation;
- academic data archives maintaining and encouraging re-use electronic resources of interest to specific academic communities;
- legal deposit or copyright libraries with a statutory obligation to maintain and provide access to non-unique information objects.

The information landscape covered by the framework is therefore rich and varied and its implementation will be tailored to the specific needs and responsibilities of individuals and institutions. However whatever the needs and responsibilities, we believe those individuals and institutions will benefit from considering the framework in developing appropriate policies and implementation guidance. In addition it is our belief that the roles of different stakeholders in long-term preservation of the cultural and intellectual heritage cannot be achieved without consideration of the life cycle of the resource and the co-ordination of the separate interests as embodied in the framework.

### 4.4 Legal and Economic Environment

Not a stage in the life cycle of a digital resource but a consideration of the legal and economic environment surrounding the resource and interlinked with the organisational mission of its stakeholders which will also impact on the life cycle and the application of the framework. Legal issues may include: intellectual and property rights in the resource or integral software supplied with it; contractual terms attached to a resource or the hardware and software needed to access it; protecting the confidentiality of individuals and institutions; protecting the integrity and reputation of data creators or other stakeholders in the resource; or any legal obligation to select and preserve the authenticity and content of categories of records or individual resources. What rights are vested in a resource will impinge on how and whether it may be represented in machine-readable form; how, by whom, and under what conditions it may be used; how it can and should be documented and even stored (e.g. where 'sensitive' information requires encryption or access restrictions); and how, whether, and by whom it can legally be preserved.

Similarly the business environment(s) in which a resource is created, managed, preserved and used will have a bearing on the application of the framework. Resources created in a commercial environment may have a commercial life cycle which can impinge on data management, preservation, and use. Some organisations may also be subject to more sudden and abrupt changes in ownership and rights, or location and data management than others.
The returns required on investment in resources may also require physical control of storage and access, and/or systems and procedures for encrypting, marking or locking the resource, user registration and authentication, charging, and rights management. All of these can affect and in some cases can mitigate against long-term preservation unless they are specifically addressed as issues and the requirements of different stakeholders can be met.

The priorities and objectives of funding, and the funding agencies, for the resource through the life cycle can also vary and impact in a number of different ways. This is particularly important for documentation and metadata on the context and content of the resource which are most easily developed or captured when the resource is created and can only be re-constructed at greater expense, if at all, at a later stage of management and preservation. The cost-effectiveness over the life cycle of the resource of completing data documentation and metadata when the resource is created (and often its immediate benefits to the data creator) needs to be recognised and its practice encouraged.

### 4.5 The Life cycle of the Resource

*1. Data creation*

Data creation will normally involve a design phase followed by an implementation phase in which the data is actually created. Consideration of the framework will have its greatest benefits during the phase of developing funding, research and project designs, design of information systems, and selection or development of software tools.

The decision to create digital resources can be undertaken for a number of different purposes and involve a range of stakeholders who will have some influence on the process. Data creation may be undertaken by those creating information from its inception in digital form (primary data creators), or by those involved in the creation of digital materials from information in traditional media (digitisers). The timescale for creation of these digital resources can be finite and definitive or dynamic and continuous. In some cases hybrid resources incorporating both digital and traditional media may be created or the resource hyper-linked to other resources.

Each of these processes and the form of resource entail a range of decisions which will involve selection and determine a data resource's cost, benefits, intellectual content, fixity, structure, format, compression, encoding, the nature and level of descriptive information, copyright and other legal and economic terms of use. Accordingly how data is created and its form will impinge directly upon how it can be managed, used, retained and preserved at any future date. All or most of these criteria will also determine a resource or collections usefulness to the data creator and funding agencies and its fitness for its intended purpose.

The process of data creation by individuals or institutions may be influenced by a number of different stakeholders. Funding agencies, publishers, and software developers can influence or determine different aspects of the decision process. Curators interested in the development of policies and guidance for the creation and long-term preservation of the resource should therefore identify strategic partnerships and dependencies and ensure that these are addressed. This will usually involve developing a dialogue with internal or external data creators, users and other stakeholders, and considering the implications of how a resource has been created and documented for its management, preservation and future use.

*2. Data and Collection Management and Preservation*

Data and collection management and preservation may involve a number of stakeholders who can fulfill different functions and roles. These functions and roles may be for a fixed or indefinite duration and can involve direct or indirect participation in the process. Immediately after creation of the data and usually for a period after this the primary data creators and digitisers will be responsible for the management and short-term preservation of the resource. The resource can also be deposited or w ill be transferred at a subsequent point to institutions or internal departments which will support or assume responsibility for long-term preservation and access. These functions can be undertaken by internal departments within the digitisers where their organisations' roles extend to long-term preservation. Alternatively these functions will be achieved by offering to deposit with and/or acquisition of the resource by the institutional archives, copyright and deposit libraries, and academic archives.

In addition, digital information may be created as part of the process of collection building or collection management of a resource. This can be seen as an extension or supplement to data

creation process and similar criteria will apply. Collections may be extended or new aggregations of resources created by licensing, copying or mirroring existing digital information created by others. New digital information can also be created in collection management processes e.g. the computerised cataloguing or digital research materials generated from existing resources in digital or traditional forms.

In some cases the resource or collections may be managed and preserved by administrative processes which we have described as "remote management".

For dynamic constantly changing information, a single deposit and acquisition for long-term preservation may be inappropriate. In such cases digital information may remain with the data creator who will assume responsibility for updating and maintaining it. The primary data creator may be legally obliged or voluntarily abide by standards and procedures established by an external organisation with established procedures for deposit. Decisions may be taken to periodically sample or copy the resource which will provide an archive of the resource at particular points in time.

"Active" resources which are still used by their creators in a current project or business process may be managed and preserved by a similar process of remote management in which the data creators abide by standards and procedures agreed with and monitored by an external organisation. In such cases the data may be reviewed and selected for deposit and acquisition when it is no longer in an active phase of use by the data creator. Alternatively a copy of the data may have been deposited during this active phase but access may be denied or restricted for an agreed period.

The organisations we have identified as "data banks", and to a more limited extent other organisational types, may also be involved as contractors in remote management of resources. They frequently manage resources under contract to others who retain legal responsibility for the resource and set terms and standards in the contract for their management.
The main processes involved in data management and preservation can include the following:

Acquisition, Retention or Disposal

Acquisition of a resource may involve decisions about collection policy, selection and rejection criteria, sampling methodology, collection levels, retention periods, disposal of part or all of a resource, selection for long-term preservation, and which data resources should be accessioned into (or excluded from) a permanent collection or handled by remote management of the resource. It will also involve data evaluation - a nuts and bolts assessment of those data resources which are potential acquisitions and will determine how (even whether), and at what cost a data resource may be included in a collection and its fitness for its intended purpose.. This process will be critically dependent on or affected by decisions made when the resources were created: the formats and structures used, data quality and consistency, the existence of metadata and documentation, or the rights accompanying the resource. Decisions taken when the resource is acquired will subsequently shape the collection and impinge directly upon how it is catalogued and documented, managed, made accessible to end users, and preserved.

The selection process occurs primarily when the resource is acquired but can be an iterative process. Decisions not to retain a resource, or to transfer it to another organisation can occur after an agreed review period or as the collection policies of an organisation and its peers evolve and change over time.

Data management

A suite of related decisions about how data resources are handled and described once they are included in a collection. How data is managed will depend upon how it has been created or supplied (e.g.. in what format, with what documentation, and under what terms and conditions). Data management options will accordingly be constrained by decisions taken when data is created or selected for inclusion in a collection and by the funding and technology available to the organisation. They will also constrain data use and preservation options. The suite of

decisions are outlined below in greater detail:

Data structure, format, compression, and encoding.

How data is formatted (written to magnetic media), compressed, and encoded (i.e. how internal semantic or syntactic features are represented) will determine its portability across hardware and software platforms and how it may be stored, manipulated, and subsequently enriched.

Data description and documentation.

The information supplied about a data resource's structure, contents, context, provenance, and history. The information will normally be in two parts; information which was created with the resource such as users' manuals and data dictionaries or provided to document its transfer; and secondly new digital information created when existing resources in traditional or digital form are catalogued or supplemented by research. It influences how a resource is located, managed, and used, and frequently reflects data acquisition decisions (notably as they reflect what documentation is supplied for a resource, how it is supplied, and who supplies it), and the subject or sectoral documentation standards and practices of the creators and curators of the resource. It will also be contingent upon the resources in terms of cataloguing staff and expertise available to the managing agency.

Data storage.

Involves organisational decisions about whether collections or parts of collections are stored centrally or distributed across several sites, contracted out to a data bank, or the technical decisions about what magnetic media and hardware platforms, physical security, refreshing or replacement of storage media, and contingency procedures, are used. Options are constrained by the resources' structure format, compression, and encoding; by whether the resource is dynamic or fixed in its nature; the need t o maintain authenticity and integrity of the resource; and also upon the relative emphasis given to their use and/or preservation. Accordingly data storage decisions together with the available funding and technologies can constrain data creation or acquisition and help to determine how (even whether) and to what extent a data resource once included in a collection can be preserved and/or used.

Data storage will involve decisions on the short-term preservation of the integrity and functionality of the resource, which will normally involve a combination of the following:

- periodic checks of completeness, function and consistency of the resource;
- refreshing the storage medium and copying the resource to overcome any instability in the medium over time;
- Migrating the resource onto new storage media or into new formats
- the provision of contingency copies with storage in multiple locations to safeguard against damage or loss;
- retaining a copy of the resource in its primary format before any migration for future checking and validation and if necessary recovery of data.

Data preservation.

A suite of strategic and procedural decisions which together with other aspects of data management help to ensure that the content, context and authenticity of a data resource survives through time and changing technologies with minimal loss in its information content, functionality, and accessibility. Decisions involve the adoption of a preservation strategy or combination of strategies normally taken from the following list:

- migration (data is stored in software-independent format and migrated through changing technical regimes);
- technology preservation (data is preserved along with the hardware and/or software on which it depends);
- emulation (the look, feel, and behaviour of a data resource is emulated on successive hardware/software generations);
- long-term preservation is highly contingent on decisions taken when the resource is

created and during its subsequent management, and also rests on available funding and technologies. It is also undertaken to maintain future access and use of the resource and is therefore closely linked and potentially contingent upon data use.

*3. Data Use*

Data use can occur immediately after its creation and for an indefinite period thereafter. Its use can be to fulfill its primary purpose when created, involve subsequent secondary analysis, or inclusion in a collection developed to fulfill other aims. The primary data creators, digitisers, funding agencies, publishers, institutional archives, copyright and deposit libraries, academic archives and their user communities may all be involved in data use or defining and servicing user requirements. Use of the data will be highly contingent on the decisions made and circumstances surrounding creation, management and preservation of the resource; the rights management and economic framework which applies, and the approaches taken to identify and reconcile the needs of different stakeholders.

How data is delivered to and used by end users will be contingent upon: how and why it was created or acquired; agreements to co-operate, share or exchange data between different institutions; conditions and procedures required to meet legal and economic requirements; how/where it is stored; and upon what software and hardware is needed to access it. It's use over extended periods of time will also be contingent on decisions made on data management and preservation.

## 5.Case studies

The applicability and use of the framework varies between organisations according to their mission as regards the creation, management, and use of digital information; to their funding; and to a certain extent upon the availability of (and organisation's access to) appropriate technologies. Organisational mission proved to be the key determinant when analysing interviewees responses to the study and the investigation revealed five roles. Individual institutions may not be confined to a single role, but normally a single role had the greatest influence on its approach to data creation, management, and use.

- **Data banks**. Data banks such as university computing services perform large-scale data storage functions for a broad constituent community. They are contract data services whose core function is to act as safety deposit boxes into which data creators deposit their data for safe keeping under some form of agreement, and from which depositors again may recall their data at some point in future. The data bank ensures that deposited data are available on contemporary magnetic media and leaves depositors to worry about whether they can be represented on and meaningfully accessed with contemporary hardware and software. In some cases, the data bank may also contract with a depositor to take on certain functions which are more closely associated with an institutional or academic data archive, though these may be said to be additions to their core services.
- **"Digitisers "**. Digitisers create data resources or build collections of resources which are either created or somehow acquired from third parties) for a variety of different but always very specific purposes. They exercise a substantial degree of control over the data creation process and their use of the framework is influenced by their focus on the particular purpose or purposes to which their data collections are to be put. It is possible to group the digitisers that were interviewed into three broad categories which reflect their roles and their intentions in the data creation process. Those categories include:
- Research-oriented agencies and individuals create or acquire data resources in the course of (or as an output from) specific investigations.
- Library, archive, and cultural heritage organisations. Such institutions have existing collections made up predominantly of non-digital information objects. Their data creation and acquisition activities are guided by collection policies which govern the institution's curatorial work generally and focus in four main areas: collection management and accountability (e.g. through the creation of computer catalogues); collection development (e.g. by acquiring access to third-party data resources as a means of appropriately extending the institution's "holdings"); access to the collections (e.g. through the creation and network delivery of digital surrogates for objects within the collection); preservation

(e.g. through the creation of digital surrogates for at-risk objects within the collection); repair (e.g. through the creation of digital surrogates for damaged objects within the collection which may guide repair). It is likely that the organisational missions of this group will develop over time as the balance of collections move towards objects in digital form and as those collections include an increasing proportion of accessions created as primary digital objects. At this point it is likely this group will increasingly resemble other groups such as academic data archives which preserve and promote access to digital resources of long-term value. The current focus on the process of digitisation and the creation of surrogates in digital form will then be less dominant.

- Publishers produce primary or secondary data for commercial purposes. No electronic publishers were interviewed in the course of this investigation.
- Funding and other agencies which invest in the creation of digital information resources and or exercise some strategic influence over the financial, business, and legal environments within which such resources are created. Positioned to determine how and why data resources are created, these agencies may have a determining role in whether, how, and at what cost, data resources will be managed over the long-term, and made accessible for re-use. Their use of the framework may help to extend their influence over data resources throughout each stage of their life course.
- Institutional archives, such as government or business archives selectively build and manage unique electronic records which are generated by an organisation and retained by that organisation to document its activities. They will also make deposited records available as required by the record-generating organisation. Institutional archives' use of the framework is governed by their involvement with unique records, their interest in those records long-term retention, their influence, through the record-generating organisation, over the behaviour of data creators, and their reliance upon mandated deposit by those creators as a source of collection development.
- "Academic" data archives. Academic data archives selectively develop, maintain, and encourage re-use of unique data resources which are of interest to particular end-using communities. The resources themselves are drawn from a wide variety of depositors, though once deposited, they typically become the curatorial responsibility of the academic data archive. The archives' use of the framework is influenced by their focus on secondary analysis, by their service to a specialist user community, by that user community's information requirements, and by their reliance upon voluntary or non-exclusive deposit as a means of collection development.
- Legal deposit or copyright libraries. Copyright libraries have a statutory obligation to maintain and provide access to non-unique information objects whose deposit is legally prescribed and enforced upon producers of certain classes of those objects. Copyright libraries may supplement these core holdings through voluntary deposit and, funding permitted, through acquisition of objects either through subscription or purchase. Their use of the framework is governed by their reliance upon mandated deposit, their lack of influence over depositors behaviour, their orientation toward long-term preservation and secondary use.
- Although the institutions involved in the study frequently combined the functions of one or more of these types, the case studies that are set out below concentrate on and represent the principal focus of their work with digital information.

### 5.1 The Data Bank

Introduction to case study

Data banks provide a core contract data service as a safety deposit box into which data creators deposit their data for safe keeping under the terms of some agreement. As with money and other valuables which are deposited in a safety deposit box at a High-Street bank, data deposited in a data bank are typically only accessible to their depositor. The analogy between the data bank and the safety deposit box may be extended still further. The bank which provides the safety deposit box is responsible for ensuring that money and other valuables are available to the depositor at any point in future. It is not responsible for ensuring that the money and other valuables have any use or value when they are withdrawn. That responsibility is left with the depositor. If a depositor fails to withdraw and exchange currency before it is rendered obsolete by geo-political or other

changes, the bank cannot be held responsible. Similarly, the data bank is responsible for ensuring that deposited data are readable on contemporary storage media. It is only responsible for ensuring that those data can be meaningfully represented on and accessed from contemporary hardware and software platforms if it is explicitly contracted to fulfill these additional functions by the depositing organisation. Otherwise, that responsibility is left with the depositor. To fulfill their core functions, data banks rely upon extensive infrastructure (large-scale computers, robotic tape libraries) and the large-scale deposits which justify their expenditure on it.

Representatives of two organisations fulfilling the core functions of a data bank were interviewed. The Oxford University Computing Service (OUCS) provides an archive for the electronic assets of the University of Oxford. The University of London Computing Centre acts as a data bank for a variety of depositors and offers a data bank facility for the UK's Public Record Office's Computer Readable Data Archive. OUCS's archiving service is offered to projects conducted by University staff which generate data deemed by the University's Computer Archiving Group - a group responsible for developing and implementing the University's archiving strategy - to be of value to the University as a whole and not just to an individual, a department, a faculty or a college. The service is offered upon application for five years or for the life of the data generating project which ever comes first, though extensions may be granted upon further application by the project to the University's Archiving Computer Group. The costs of archiving a project's data during the first archive period are met by the OUCS from its annual operating budget. Where the archive service is extended beyond the initial period, the OUCS may seek to recover the costs of archiving from the data generating project. In certain exceptional circumstances, the University's Computer Archiving Group may identify a data resource as an essential information asset of the University and take over responsibility for acquiring and allocating funding to the OUCS as required to ensure the data's long-term preservation.

The ULCC acts under contract to the UK's Public Record Office (PRO) as the repository for some of the electronic records and information systems created by UK government departments and selected for long-term retention by the PRO. Like OUCS, the ULCC is principally responsible for preserving archived data at the bit-stream level. Additionally the ULCC is contracted by the PRO to distribute those data physically to secondary users (i.e. by transferring them on some magnetic media or via ftp) and to make at least some of them accessible online. In these respects its involvement with PRO data takes on some of the characteristics associated with an institutional archive as described in greater detail below.

Data creation and collection development

*Data creation*

Acting on a contract basis to manage data at the bit-stream level, with no interest in a data resource's future usage, and compelled for economic advantage to offer the same service to all , the data bank has little interest in how, why, or for who m deposited data are created. This unique perspective is apparent in the core services offered at both OUCS and ULCC. Both organisations accession and store data created in a variety of different standard and non-standard formats. It should be noted, that as a university computing service, OUCS also acts in a demand-driven advisory capacity offering guidance to data creators and would-be depositors about data formats and documentation which may be more or less appropriate for the purposes of their short- and long-term preservation. Where ULCC's work with the PRO is concerned, PRO guide-lines pertaining to the management of computer-readable datasets mitigate to a large extent the need for that role being taken up by the data bank.

Data acquisition

The data bank operates on a cost for quantity economic model. Accordingly, its role in data selection is limited. Having said that, both the OUCS and the ULCC depart some way from the ideal type data bank, OUCS because it is represented on the University's Computer Archiving Group. ULCC departs from the ideal type in its work for t he PRO. Although the ULCC must archive all data resources and information systems deposited by the PRO, it does exercise some influence, in discussion with the PRO about accessioning priorities and costs, and with officials

in government departments who a re responsible for identifying and preparing records for long-term retention.

*Data management*

Data structures and data storage

The data banks leave responsibility for how data are formatted, encoded and compressed, with depositors, though may regulate how (e.g. on what media) deposited data may be transferred. Data banks are therefore largely unconstrained in the data structures they can cater to and will not normally need to restructure data unless they are contracted by the depositor to perform content migration or data distribution functions or to provide access services. OUCS will undertake these additional functions w hen engaged (and funded) to do so either by the record-generating project, or by the designated University authority which may take responsibility for the long-term preservation of certain data resources. The ULCC acts in a similar capacity, though with a smaller range of data formats than are likely to be deposited at OUCS at least with regard to its PRO-deposited holdings. This is due to the fact that government departments take account of data resources' physical and technical characteristics when selecting data for deposit with ULCC, a subject which is taken up in greater detail below. ULCC will also restructure data deposited by the PRO since it is engaged by the PRO to migrate them through changing technical regimes and make them accessible to users . Again, more on these subjects below.

*Data description and documentation*

With the exception of essential administrative information which is supplied by the data bank to locate, name, and record other vital statistics about deposited data, data documentation is left entirely to the depositor. Again, ULCC's role is exceptional where PRO data are concerned, since the PRO has contracted out to it some functions in standardising and enriching documentation that is supplied by depositors.

*Data preservation*

Data banks migrate data files through storage media to ensure their readability. Content migration (ensuring that data can be meaningfully represented by and accessed from contemporary platforms) is the responsibility of the depositor. The data ban k will rely upon extensive computing infrastructure which may include large-scale computer servers, robotic tape libraries, etc. Preservation is based around the management of archive copies of the deposited data resources; that is, copies which are independent of any on-line representation they may have. The following preservation scenario is an ideal type compiled from interviews with representatives of those institutions which provide full warehousing facilities and may not exactly reflect procedures used at any one of those institutions.

- Archive copies are stored on industry standard digital tape or other approved media as may arise, and there will be multiple copies of any single data file, some stored on and others stored off site, preferably in temperature controlled and fire-proof safes or rooms. Off-site copies should be a safe distance from on-site copies to ensure they are unaffected by any natural or man-made disaster affecting the on-site copies.
- Archive copies may be written with different software to protect data against corruption from malfunctioning or virus- or bug-ridden software, and may be made to comparable magnetic media purchased from different suppliers to guard against faults introduced by the media's suppliers into their products or into batches of their products.
- Data files stored as archive copies will be migrated periodically to new media with that migration taking place within a minimum time which reflects the media supplier's estimate for the media's viability under prevailing climatic conditions. In addition, media will be checked periodically for their readability. Such checking may be conducted automatically by archive systems according to parameters set by system operators.
- The integrity of data files may also be checked using checksum and other like procedures which may be implemented automatically by the archive system according to parameters set by system operators.

*Data use*

Beyond ensuring that depositors can recall their data on readable media, the data bank is unconcerned with data's re-use, although ULCC's position is complicated by its having been contracted to the PRO to distribute holdings in its Computer Readable Data Archive, and in this respect, to adopt functions more typically associated with an institutional or academic data archive.

User support is oriented exclusively toward depositors (typically also the data's sole users) and may include documentation about the service on offer, how it works, and how access to it may be acquired. At OUCS, the lion's share of this documentation is available on line. At ULCC, user support services are complicated by the data bank's involvement providing third-party access to PRO-deposited data.

*Rights management*

Since the data depositor tends to be the sole user of data which are stored in a data bank, rights management is not a central concern. At OUCS, depositors take full responsibility for data they deposit in the archive and there are some indemnities protecting the OUCS against any claims that might otherwise be made.

**5.2 The Digitisers**

*Introduction to case study*

Digitisers create data resources or build collections of resources which are either created or somehow acquired from third parties for a variety of different but always very specific purposes. They exercise a substantial degree of control over the data creation process and their use of the framework is influenced by their focus on the particular purpose(s) to which their data are to be put. It is possible to group the digitisers into three broad categories which reflect their roles and their intent ions in the data creation process.

- Research-oriented agencies and individuals create or acquire data resources in the course of (or as an output from) specific investigations. In many cases, the data will be primary materials for which no non-digital analogue exists, for example , remote sensing data or statistical and other databases. In others, electronic texts created for linguistic or stylistic analysis, the data will be created as a surrogate for an underlying source.
- Library and cultural heritage organisations. Such institutions have existing collections made up predominantly of non-digital information objects. Their data creation and acquisition activities are guided by collection policies which govern the institution's curatorial work generally and focus in four main areas: collection management and accountability (e.g. through the creation of computer catalogues); collection development (e.g. by acquiring access to third-party data resources as a means o f appropriately extending the institution's "holdings"); access to the collections (e.g. through the creation and network delivery of digital surrogates for objects within the collection); preservation (e.g. through the creation of digital surrogates for at-risk objects within the collection); repair (e.g. through the creation of digital surrogates for damaged objects within the collection which may guide repair). It is likely that the organisational missions of this group will develop over time as the balance of collections move towards objects in digital form and as those collections include an increasing proportion of accessions created as primary digital objects. At this point it is likely this group will increasingly resemble other groups such as academic data archives which preserve and promote access to digital resources of long-term value. The current focus on the process of digitisation and the creation of surrogates in digital form will then be less dominant.
- Publishers produce primary or secondary data for commercial purposes. No electronic publishers were interviewed in the course of this investigation.

The institutions selected for interview were involved in one or several of these activities. Research organisations included the Space Data Centre at the Rutherford Appleton Laboratory (SDC). Discussion about such organisations is supplemented with information taken from a

variety of secondary resources pertaining to data creation methods and approaches adopted by researchers in various disciplines. Cultural heritage organisations and libraries include the British Film Institute (BFI), the National Museum of Science and Industry which includes the Science Museum in London, the National Railway Museum in York, and the National Museum of Photography, Film and Television in Bradford (Science Museum), the University of Michigan Library (UML), and the Victoria and Albert Museum (V&A). Although respondents selected for interview represent a reasonable range of data creators, further work is recommended, notably amongst publishers who act as primary or surrogate data creators but in an explicitly commercial context.

*Data creation and collection development*

Data creation and acquisition

The digitisers exert maximum influence over the data creation process and do so primarily to ensure that data resources suit the purposes for which they are intended and which are outlined above. The digitisers interviewed for this study recognised how content and technical decisions taken by them impinged upon whether, how, and at what cost data once created would serve the purpose(s) for which they were intended and be managed in future. With these objects in view, they paid close attention to appropriate data standards, evaluating existing ones and adopting those which proved most appropriate. Standards typically considered may be grouped as follows:

- Technical standards to facilitate data resources' interchange across networks and between platforms with minimal loss in content and functionality. Such standards include those pertaining to file formats, and compression and encoding techniques.
- Data documentation standards to facilitate data resources' management and meaningful interchange between individuals and organisations.
- Controlled vocabularies and other standards which help to ensure that data resources are comparable with other like resources. Amongst surrogate data creators, such standards were used almost exclusively in describing or documenting their data resources. Amongst the primary data creators, they were used more widely, for example, to supply normalised or standardised values for categories of spatial or prosopographical data.

The digitisers paid as close attention to best practices defined here as the constellation of technical, documentation, and data standards and of methodological implementation strategies which promise to maximise a resource's intended usefulness while minimising the cost of its creation and subsequent management and use.

In all cases, the range of standards and best practices that were evaluated and then selected by the digitisers was contingent upon the data type of the resource being considered (file formats appropriate for electronic texts are different than those f or digital images), upon their ability to support a data resource in its intended use (the file format appropriate for Web-accessible image thumbnails may be different than that appropriate for digital surrogates created for the purposes of preservation), upon their cost of implementation and future maintenance, and upon available technology, in that order of importance. Where documentation standards were concerned, intended users' information requirements, and professional curatorial or specialist practice within the digitising institution (library professionals inclined toward MARC and MARC crosswalks; museum professionals towards other standards such as SPECTRUM, research organisations toward practices known amongst specialist researching communities), were also influential in standards evaluation and selection.

The standards and best practices which promised best to facilitate and reduce the cost of a data resource's long-term preservation were not always those which promised best to facilitate and reduce the cost of its intended use. The standards and best practices which promised to ensure a data resource's maximum fitness for purpose were also not always affordable or technically achievable. Accordingly, the selection of standards and best practice frequently involved a range of compromises between data creation aims and costs.

The digitisers carefully evaluated the data resources they proposed to create or acquire access to and used both content and technical criteria in the process. Research organisations tended to

create or acquire only those data needed for immediate analytical use. The SDC, for example, hosts a range of projects which involve data-generating space craft. Although the SDC acquires all such data generated from its space craft, some processing and selection is involved owing to the constraints imposed by the sheer volume of data generated and the costs involved in their maintenance. The SDC will also acquire research data from third parties but rarely act as the primary archive for such data. For example in the SOHO project, the receiving Centre and primary archive is the Goddard Space Center in the US which distributes a duplicate dataset to SDC for further research.

Where based within institutions with existing collections, the content criteria applied by the digitisers to the acquisition of data reflected the institutions' collection development policies. Thus, UML the BFI, the V&A, and the Science museum all created digital surrogates of objects within their collections in order to extend public access to them. The UML and the V&A through its library also acquired access to third party content in order to extend their collections. In all cases, data resources created or acquired from third parties were evaluated in terms of their fit with users' current information requirements; their ability to fill gaps in extant holdings; and the cost of their creation, acquisition and maintenance (particularly import ant where digital versions of extant information objects were being considered). Data resources intended as preservation surrogates were evaluated with regard to the wear and tear and other threats to the integrity of the underlying information object.

Technical criteria which were used in the evaluation process tended to be of more recent origin than any institutional collection development policy and had been established to take account explicitly of digital information as created or acquired. They were used for three purposes: to determine data resources' fitness for purpose; to determine the nature and cost of the technical and infrastructural requirements to create or acquire access to them; and to determine the nature and cost of the technical and infrastructural requirements to support their use and management. In all cases, evaluation criteria reflected organisational mission and in this respect were strategic. The UML created or acquired access to digital objects to extend, enhance access to, and preserve its holdings. Amongst the museums interviewed, third -party data resources were a low priority for the primary collections but were more commonly used in developing the museum library. Data creation was seen first and foremost as a means of collection management and accountability (e.g. through the construction of comprehensive computer catalogues), and thereafter to extend public access to those holdings. Digitisation for access was particularly compelling where public access was severely restricted (as for example, with the BFI's film collections which are only accessible to audiences able to attend scheduled viewings at selected regional theatres). Digitsation for preservation was also apparent in the museum sector and involved criteria similar to those used in the library (the BFI again has created digital surrogates for analogue recordings of interviews with key personalities in the film and television industries which are stored on volatile physical media). There was also an interest in digitisation for repair. The National Gallery for, example, trial s repairs on digital surrogates of paintings before implementing those repairs on the paintings themselves.

Both the content and technical criteria adopted by the digitisers were also opportunistic and in this respect sensitive to available funding, technologies, and, where digital surrogates were concerned, to available underlying content. The Science Museum is in the process of developing a digital image policy which includes cost-benefit analysis of storage, access, and image quality options to guide staff in selecting appropriate standards and formats for image content which is being considered for digitisation. Elsewhere, funding and technology had other obvious impacts. The UML oriented a significant share of its digitisation effort around Americana in response to a grant from the Carnegie Trust. The BFI is digitising 30 hours of film of academic interest in history, medicine, and performing arts, reflecting a fruitful partnership with the Joint Information Systems Committee of the UK's Higher Education Funding Councils. Technical criteria may be as constrained. The BFI creates digital surrogates for volatile sound recordings because it is technological possibly to do so without any appreciable deterioration in sound quality. Further, such digital reproductions can, with some constraints, be delivered to end-users via the network and other means. Similar "archive quality" reproductions are not possible to create or deliver where digital film and video is concerned ruling out digitisation for preservation of film and video. The availability of underlying content finally is a consideration. The BFI's selection of

film and video for use in the BFI/JISC project is restricted to content which is free and clear of any copyright considerations.

The digitisers's primary focus on digitisation for access entail an interest in de-selection of some digital information objects within their collections. In research organisations, a data resource may be destroyed, or deposited or returned to that agency which takes primary curatorial responsibility for it upon completion of the investigation with which that resource is associated that resource. In cultural heritage and library organisations, de-selection will conform to collection development and management policies. It may take place when a data resource is superseded by a new version, for example, where a digital library mounts successive versions of a bibliographic database supplied with regular updates and amendments by some third party. It may take place where a cultural heritage or library organisation creates a duplicate surrogate with greater functionality than that already held.

*Data management*

Data structures and data storage

How data were structured and stored reflected their nature (digital images were treated differently than electronic texts and catalogue records) and their intended use, but also the complex mix of organisational, funding, and technological constraints which drove content selection decisions. In some cases, copies of the data are stored in different structures, typically where a master copy of the data resource is used to generate subsequent copies structured as appropriate for different delivery an d use scenarios.
In the research organisations, data structures reflect the needs of the research communities by or for which they are being assembled and accordingly are highly contingent upon the methods and needs of those communities. In some cases, such data will be stored in a proprietary format as required by particular application software. With regard to storage, data are stored for the life time of the research project with which they are affiliated, although the SDC did provide long-term storage facilities for some of the data generated by space craft affiliated with its own projects. Such data are typically stored on line so long as they are being used actively in research, but may be moved off line, deposited or returned to an academic data archive, or destroyed, when active research ceases.

In the library and cultural heritage organisations, data storage and data structures also reflect the kinds of data being processed and the reasons for which they were assembled. With regard to the construction of catalogues, data storage was centralised in support of network access, although the BFI and the UML were considering tools to integrate access to distributed catalogues whether maintained in-house or by multiple third parties. With catalogue data structures, the fields or elements of information supplied in each record were of far greater concern than the file formats in which such records were stored (the latter were typically available as field delimited ASCII files). Selection reflected users' information requirements but also the nature o f the information object being described in the catalogue record. The catalogue record used by the BFI to describe its film digital (and non-digital holdings) comprises very different elements than that used by the UML to describe its electronic texts. In every case, the digitisers evaluated and, where possible, adopted appropriate cataloguing standards, only resorting to proprietary practices as a last resort. The UML used a slightly modified version of the Text Encoding Initiative's recommended SGML header. The Science Museum conformed to the Museum Documentation Association's SPECTRUM standard. The V&A uses MARC for its library catalogue records and conforms to the SPECTRUM standard for its collection information system. The BFI alone developed its own standards but only because it began to develop its catalogue of holdings before any standard practices had emerged for complementary collections. In so doing, it helped to established a *de facto* cataloguing standard for film and video collections.

Where information objects are digitised for the purposes of access, users' needs, costs, and available technologies are once more paramount in determining data storage and data structures. For its electronic texts, the UML binds two digital representation of the same underlying text - an SGML-encoded ASCII transcript and a raster image - in order to enhance functionality. The text transcript facilitates keyword searching and other more refined searches. The raster image

provides user access which may be more appropriate for reading or printing the text or for checking the accuracy of the text transcript. The strategy increases the UML's data creation costs, though marginally and not enough to off-set perceived gains in functionality. Still economies are sought and text transcripts are only lightly marked up - that is, SGML tags are used to identify a small number of textual elements.

With the BFI's the same considerations exercised a different influence over the digital film projects. The BFI operates two such projects. The one comprising 30 hours of content for use by UK higher education has already been discussed. The other provides a further 30 hours of digital film to a broader end user community. Because UK higher education has access to broadband network technologies and large regional computing servers, and because users have access to relatively high-end desktop computers, the BFI was able to create and distribute high-quality digital film and video reproductions for storage at and distribution from regional computing centres. Given the greater network constraints involved in this second pilot, and the likelihood that end users would have access from a range of desktop computers not all with high end specifications, it chose to compress the digital surrogates to a far greater extent (and accordingly to sacrifice some of the reproduction quality), and to distribute it from a central site (maintained by the BFI).

The Science Museum which has created some 2,000 digital images created to provide access to aspects of its collection, has selected PhotoCD and converts images to JPEG for the purposes of access. The V&A follows an identical procedure for the 7,000 -8,000 digital surrogates it has made from the 60,000-80,000 slides in its slide library. In both cases, PhotoCD was selected because it is widely used, cost effective and easy to implement, and, although a proprietary format it supports output to a range of formats suitable for access purposes.
Where digital information objects are acquired from third-party suppliers as was the case at the UML, the digitisers may seek directly or indirectly to exercise control over how third-party data are structured. UML, for example, will not normally purchase or subscribe to those data resources which cannot affordably be integrated with other data resources in its collection.
Where digitisation for preservation is concerned, end-users' perceived needs are less evident than the digitiser's interest in ensuring that the surrogate represents the underlying information object with as little content loss as possible. A common question applied by those interviewed was, "what information would be lost if the surrogate was the only surviving example of the underlying information object?" Here, the digitisers rely upon their evaluation of the most current research and on best practices where they are evident. Thus, the UML investigated recommendations made by Cornell University about image qualities appropriate for digital surrogates of printed books. Similarly, the BFI followed industry standards for digital sound reproductions wit h regard to its audio surrogates. Preservation surrogates were in all cases stored centrally in house.

Data restructuring

In the research organisation, any data restructuring that occurred did not take place to facilitate access or preservation, so much as a result of the normal processing involved in the analysing the data. It was not, in this case, purposeful. In the library and cultural heritage organisations, restructuring was purposeful, only there the digitisers sought to minimise the extent to which it had to take place owing to the expense involved. All recognised that reformatting may become necessary to migrate data resources through changing technological regimes for the purposes of their preservation. Indeed, most had taken this possibility into account when determining how data resources were created in the first place, selecting file formats and compress ion techniques which were widely used, even if they were proprietary.

In some cases involving digital surrogates created to enhance access, the digitisers in libraries and cultural heritage organisations accepted that restructuring could improve the usefulness or functionality of a data resource and had accordingly adopted or were considering strategies for periodic restructuring. The Science Museum converts PhotoCD images to JPEG for delivery and is in the process of evaluating the various JPEG compression algorithms which will facilitate and speed the network delivery of these images. As already discussed, the UML may reformat and restructure digital resources acquired from third-party suppliers in order to enhance their

functionality on Michigan platforms. It is also considering how electronic texts created locally might be dynamically updated with richer encoding in response to any user demand which might materialise for more search and retrieve functionality. The BFI expects that digital film and video will be compressed to varying degrees for delivery to users accessing digital collections from different computer hardware and network environments.

Data documentation and description

At least two levels of documentation were apparent for the digital objects created or acquired by the digitisers: one richly descriptive of the data resource's content, structure, and provenance, and another, an abbreviated subset of the former as appropriate for entry into an on-line catalogue or directory of images through which users could locate and then order or gain access to the resource in question. With regard to what information was recorded about a resource, the digitisers tried where possible to conform to appropriate cataloguing standards. Within the research organisations, these standards tended to be highly specific to the information requirements of the relevant specialist community and to the kinds of data with which they were typically dealing. With regard to the documentation effort itself, data created by digitisers in libraries and museums were documented by appropriate staff. Where such data were acquired from third parties, the digitisers imposed documentation requirements up on the suppliers and then verified the documentation that was ultimately supplied by them.

Data preservation

Although the digitisers had given some thought to the long-term preservation of their data collections, it was not a central or even a pressing concern, even where digital resources were created as surrogate copies for the purposes of preservation. Several possible reasons may be adduced for this.

Firstly, the digitisers' did not necessarily perceive their data's long-term value. Amongst research organisations, for example, a data resource's value may diminish upon completion of the research project with which it is associated. At least, from the data creator's perspective, the completion of the research project marks something of a watershed after which a data resource's maintenance cost may begin to outweigh its immediate benefits. It should be noted, that this short-term assessment of research data's value is not shared by funding agencies which support their development, or by academic data archives which actively seek to acquire such data for long-term preservation and re-use.

Cultural heritage organisations and libraries may also perceive a relatively short-term value in the data resources they create. Where digital surrogates are created for the purposes of repair, for example, the repair is ultimately worked upon the underlying information object. Thereafter, the value of the data resource may diminish rapidly. Where surrogates are created for access the very existence within the collection of the non-digital sources from which the surrogates are made may diminish the data long-term value. Moreover, the focus on access naturally entails the development of on-line data stores, and the implementation of periodic back-up procedures intended to protect against unintended data loss or corruption. In some cases, the back-up procedures are deemed sufficient for the purposes of long-term. Where digital surrogates are produced for the purposes of preservation, the non-digital sources for the surrogates will frequently still reside within the collection. Even where it does not, it is rarely unique. Accordingly, the digitisers can rest assured that copies exist somewhere, even if not within their organisation's collection. Finally, where access is acquired to third-party data resources, the organisation tends to assume that the responsibility for the data's preservation rests with that third party.

Secondly, in many instances within cultural heritage and library organisations, the creation, management, and use of data resources make up a relatively small part of the organisations' overall concerns. Library and museum collections, for example, primarily consist of non-digital information objects which accordingly are the focus of their curatorial activities. Data creation and acquisition activities are still relatively recent strategic initiatives undertaken only where limited funding permits and then very much as pilot or test projects. In the conduct of such

projects, the cultural heritage and library institutions have developed an awareness of the growing contribution that digital resources are likely to make to their collections and accordingly of the need in future to implement appropriate preservation strategies. These strategies, however, are still very much at an early stage in their development.

Finally, and in a related vein, digitisers are not typically based at institutions with sophisticated and large-scale computing services with the infrastructure and expertise appropriate to long-term data management. This is especially true in library and cultural heritage organisations and appropriately reflects their historic and primary focus on the management of non-digital information. Amongst such organisations interviewed, the UML was exceptional and had in place data archiving practices which w ere roughly equivalent to those implemented by the institutional archives and the data banks.

Where the digitisers had thought about preservation, they adopted different strategies which reflected the provenance and structure of their data, the purposes for which they were created, and the computing infrastructure, expertise, and funding available locally.
Where data resources are supplied by third-parties, the digitisers relied upon those third parties (in whole or in part) for preservation. Where data resources are created in-house, the digitisers prefer migration though few have experience with it. They also identify certain data resources which require different preservation strategies. The Science Museum has collections of computer software and computer hardware. For the latter, technology preservation and emulation are required and the Museum is involved with other bodies such as the British Computer Society and the Computer Conservation Society in pursuing these aims. A collection of video games at the BFI will also require technical preservation (preferably) or emulation since the look and feel o f a video game, including the look and feel of the platform on which it was mounted) was considered a crucial part of the experience of playing the game.

Preservation practices

These vary considerably and reflect the technical infrastructure available to the digitisers and their different levels of awareness of good preservation practice. Where digitisers adopted technical standards conducive to data interchange across platforms, hardware and software issues are not-problematic. In most cases, data are stored on-line (reflecting their network accessibility to end-users), and regular back-ups are taken. Less frequently the digitisers created archive copies of their data on tape or CD - for on- and off-site storage. Those that do, have mechanisms in place for periodically checking the integrity and readability of off-line archive copies and for migrating them periodically through new magnetic media.

Remote management

Remote management is particularly evident amongst libraries which extend their collections by acquiring access for users to on-line data resources that are maintained by and accessed from third parties suppliers. The principal challenges they confront in doing so are twofold and only just beginning to receive attention: integrating access to information about third party holdings with that available for holdings on site, and ensuring access to third party resources over the longer term. A number of strategies for integrating access on- and off-site data resources are. Web-based gateways providing logically ordered or structured lists of pointers to electronic resources whether managed on- or off-site are available at the University of Virginia and help to give access to its virtual library of electronic "holdings" some of which are maintained off site. The method does not integrate information about data resources with that pertaining to non-digital objects within the collection. The Web-based gate way is not, for example, accessible to queries progressed against the library's on-line catalogue. A more integrating strategy for resource discovery permits users from a single interface to progress a single search against multiple catalogues (whether managed on or off site) and retrieve an integrated result set. An example is available from the University of California at Berkeley. Further integration which permits users to discover and then to access information objects through a single interface, are in their embryonic developmental stages, for example, at the RLG and through the UK's AGORA hybrid library project.

Safeguarding long-term access to third-party data resources was in some respects a more significant problem. Where such access is acquired through annual subscription, it is threatened where the subscription is allowed to lapse or where the third-party supplier ceases to offer the same service. How different with paper-based journals, for example. A library subscribing to a periodical for ten years is left with ten years' worth of back issues even after its subscription is cancelled or its publisher ceases to trade. The UML mitigates the problem by physically acquiring such data resources and mounting them on site. Elsewhere, the problem is being addressed through licensing arrangements which may permit libraries to obtain, manage, and provide limited access to third-party supplied digital content to which they once subscribed, even after their subscription lapses or the content is no longer available from the third party. The CEDARS project, another UK-based initiative will be investigating these issues in the coming few years.

*Data use*

The digitisers work principally with data resources intended for immediate use by definable user communities. Facilitating the user communities' access to the data resources is accordingly a principal concern and is taken account of very early on in the data resource's life cycle, for example, when its creation or acquisition is being investigated. Amongst the research organisations, use is intended for the data creators themselves or for a small and specialist community of interested specialists. Accordingly, the data are developed for use on platforms well known amongst that specialist community. Given their use by specialist and expert communities, data resources produced by the research organisations are only minimally documented and may offer little in the way of user support.

Within the library and cultural heritage organisations data resources are typically intended for use by a far broader public. Unsurprisingly they seek to exploit network technologies and the World Wide Web, although at the Science Museum, public access to some machine-readable information is mediated by museum staff. Information about the data resources is generally disclosed through network accessible catalogues, and supplied to users via the network and Web interfaces. Where the BFI's digital film an d video collections are concerned, different delivery scenarios are implemented to meet the needs of very different user communities as described above.

User support is deemed by the digitisers within the cultural heritage and library sectors as an essential means of facilitating the development and use of their data collections while minimising impact on staff time. Support for data users reflects the orientation toward on-line Internet and Web-based delivery of digital content and is itself typically web-based. Although the supporting materials vary depending upon the resource in question, they tend to contain descriptive information about the data ( including information about their content and how and why they may be used), instructional guidance with regard to the actual use of a particular data resource, and, not insignificantly, information about the terms and conditions of use of a particular data resource. Printed user support is not preferred and in some cases deemed inappropriate. Those cultural heritage and library organisation which acquire access to third-party data also offer support to potential suppliers or donors in the form of guide-lines setting out preferred approaches to the preparation, documentation and delivery of third-party data. Such documentation tends to exist in both printed and electronic formats and at some considerable level of technical detail.

*Rights management*

Rights management considerations influence the digitisers' content selection and distribution decisions.. Where library and cultural heritage organisations create surrogates of objects within their own collections, they show a preference for objects which are either free and clear of copyright or for which copyright is vested in the organisation. In some cases, for example, the UML's Old English Dictionary, or the digitised slide libraries created by museums, content created by the institution was seen as a source of potential revenue, particularly where access to it could be sold or leased to subscribers.

Where digital content is supplied by third parties, the digitisers negotiate terms and conditions of use with those suppliers on an *ad hoc* basis. In all cases, they enter negotiations with a clear understanding of what terms and conditions they (and their user communities) will accept. Thus, the UML seeks to ensure that all acquired content be made available to university members and to non-members who use the library on a walk-in basis. The Science Museum has licenses for use of digital images in its Photo library based on British Association of Photo Libraries and Archives guidelines.

Rights management has a parallel influence over distribution scenarios. Nearly all of the digitisers located in library and cultural heritage organisations have drafted user agreements emphasising personal, educational, and non-commercial use, and, in the case of third-party supplied data resources, reflecting, the terms and conditions under which access to those resources was originally acquired. They are also investigating if not immediately implementing authentication and fee transaction processing mechanisms as a means, respectively of protecting and generating revenue from at least some of the data resources in their collections.
Some of the digitisers (e.g. at UML and the BFI), express an interest in the development of outline or model data acquisition and distribution agreements which might be used to guide their negotiations with suppliers and, indeed, with users. They are not, however, sanguine about the likelihood that such model agreements will emerge in the near future or even be all that useful since such negotiations were in most cases highly contingent upon distinctively local variables.

### 5.3 Funding and other agencies

*Introduction to case study*

Such agencies invest in the creation of digital information resources and/or exercise some strategic influence over the funding, business, and legal environments within which digital resources are created. Accordingly, they are in a position to determine how and why data resources are created, and the prospects for their long-term management and re-use.
Representatives of two organisations were interviewed, the Natural Environment Research Council (NERC) and the Scottish Cultural Resources Access Network (SCRAN). The discussion also draws upon the a study conducted on behalf of the British Library/Joint Information Systems Committee's Digital Archiving Working Group on the needs of universities and funding agencies as pertaining to digital preservation.

The NERC invests in data-producing scientific research and thus in the creation of data resources which are unique, expensive to create, difficult to reproduce, and of substantial scholarly re-use value. Recognising that its investments in data-producing research may be maximised by guarding the longevity of the data and by encouraging their re-use, NERC has developed a data policy which, with the aid of high-level institutional and financial commitment, governs the disposition of NERC-funded data and acts to ensure their availability over the longer term. It has also designated a range of data centres (which receive funding directly from the NERC) which act as repositories (academic data archives) for data created with NERC funding.

The SCRAN is something of a hybrid case with characteristics typical of digitisers in cultural heritage and library institutions and of funding agencies. With money from the Millennium Fund of the UK's National Heritage Lottery Commission it is building a collection of digital information objects pertaining to the human history and material culture of Scotland. The collection is developed by cultural heritage and other organisations who are successful in applying for grant funding from SCRAN to create machine-readable catalogue records about objects within their collections and/or to create digital images of some of those objects. By the year 2000, the collection is expected to contain a corpus 1,500,000 comparable catalogue (text records) records describing Scottish cultural heritage objects which will enhance and integrate access to information about those objects. It will also contain some 100,000 multimedia data resources (most of them captioned images) representing some of those objects.

*Data creation and collection development*

The funding agencies use their money, and the application process through which it is

distributed, to influence how and why data are created and to determine their subsequent disposition and use. Both the NERC and the SCRAN have adopted data policies, informally in the case of the SCRAN, which determine the life-course of grant-funded data from their inception, through to their creation, management, and subsequent use. With regard to inception, the rigorous evaluation of both content and technical criteria that is conducted by other digitisers when planning a data creation initiative is expected by NERC and SCRAN of their grant applicants. The funding agencies are then positioned to fund only those which promise data which are (a) fit for the purpose for which they are intended, (b) created according to appropriate standards and best practices, (c) useful and re-usable and (d) manageable over the longer term. The NERC goes some way further. As added insurance against its data resources' futures, it may require successful grant applicants to work closely with an appropriate data centre in the creation of the data resources so as to ensure that such resources can be managed by that centre over the longer term.

The extent to which the funding agencies prescribe content and technical criteria varies in part owing to the range and nature of the data resources they are interested in funding. NERC is less prescriptive than SCRAN. Funding research in a wide range of scientific disciplines it has to be. Content and technical criteria which may be applied appropriately to an archaeological GIS are different than those which may be applied to satellite data.

SCRAN's criteria are more narrowly defined: content criteria by the distinctively Scottish orientation of its mission; technical criteria by the SCRAN's goal of making text records and multimedia objects widely available on-line via Web browsers. SCRAN 's criteria also reflect its interest in encouraging applications from organisations and individuals who own or manage diverse Scottish cultural heritage objects and who possess varying degrees of computing expertise and equipment. Technical criteria are particularly affected. For its basic text or catalogue records, for example, SCRAN has defined a generic record structure which can be applied to diverse collections many of which have distinctive cataloguing procedures in place. For its digital images, it prefers PhotoCD, a proprietary format produced by technology which is widely available, affordable, and relatively easy to use. Digital images created with PhotoCD can also be rendered into JPEG format for Web-based delivery. For content suppliers with access to more sophisticated technologies, the SCRAN will also accept uncompressed TIFF files as these may be manipulated to emulate PhotoCD formats and functionality. Applicants for SCRAN funding must, as a condition of grant, conform to these standards.

Given their interest in data resources over their entire life cycle, the adoption of standards and best practices are possibly even more important to the funding agencies than to the digitisers. These are considered in the development of funding agency data policies and in the evaluation of grant applications. Because it funds the development of a wide range of data resources which are created for very different purposes, NERC is once more less prescriptive than the SCRAN and relies upon specialists involved in the application review process to advise about appropriate use of standards and best practice. SCRAN's data standards are more prescriptive, though are none the less developed through consultation with appropriate communities - its generic catalogue record approximates the Dublin Core - in light of the uses to which data are intended to be put, and of available technologies.

*Data management*

Both the SCRAN and the NERC take a serious interest in how data are managed and preserved: NERC because it recognises the long-term scholarly value of the research resources created by its grantholders, and SCRAN because the digital objects created and supplied by its grantholders contribute to its on-line collections. The data standards which are imposed by the SCRAN on its grant applicants, and the technical criteria used to evaluate grant applications to the NERC both reflect these data management aims.

Again, the NERC is less prescriptive. Decisions about how to store, document, and preserve grant-funded data are contingent upon the nature of the data and upon the practices of the NERC data centre or institution where they are ultimately managed. Broadly, documentation standards which the NERC may require from its grant holders are designed and implemented by the data

centres with secondary user's information needs in view. The documentation pays particular attention to users' needs to assess quickly whether a data resource is appropriate for any analysis they intend. With regard to data storage data will typically be managed in the format in which they were created. Where restructuring takes place, it does so to facilitate preservation or improved user access. With regard to preservation procedures, the NERC relies upon its data centres which are well (though differently) equipped and which implement procedures akin to those apparent at the data banks and the institutional archives.

At SCRAN, documentation is supplied by grantholders and must conform to a the minimum level standard prescribed by the SCRAN. The documentation standard reflects the SCRAN's aims of integrating on-line access to text records which describe Scottish cultural heritage objects and to digital surrogates of some of those objects. Where the digital surrogates are concerned, a richer level of documentation is also required by the SCRAN, reflecting its interest in providing at least some minimum information about the digital surrogate's provenance and significance. Catalogue records are stored centrally by the SCRAN in an appropriate DBMS. Digital images, on the other hand, are stored in at least three, versions: a master copy or high-resolution image from which others are derived; a JPEG thumbnail or low resolution image for unrestricted on-line access; and a medium resolution educational viewing image available for on-line access to SCRAN subscribers. SCRAN accepts the master copy from its content suppliers and restructures it to produce the thumbnail and the educational viewing copies, respectively. For preservation, the SCRAN prefers migration but in practice relies upon regular back-up for its on-line holdings. Where images are concerned, the master or high-resolution copies are also the preservation copies, and two are made, one of which is maintained by the SCRAN and the other by the individual or organisation which supplied the images.

*Data use*

Here, the funding agencies may have two roles: a specific one which entails encouraging the use or re-use of the data resources created by their grant-holders; and a general one which entails promoting awareness of the scholarly and other advantage s which may accrue from the collection, professional management, and re-use of such resources generally. For NERC, both roles are undertaken through the data centres though in conformance with the NERC's data policy which is written in part with reference to the transforming effect that the availability and use of high-quality data resources may have on research into aspects of the natural environment. Practices vary across the data centres in reflection of their diverse holdings, and the different specialist user communities that each centre serves. In these respects, the data centres approximate the academic data archives which are discussed below. Across the centres, data are distributed or made accessible to users through a variety of means which include the Internet and a range of portable magnetic media, with a range of supporting materials which reflect the information requirements of the data centres' respective specialist communities, and the kinds of data which are being supplied.

As a relatively new organisation only just starting out in the development of its on-line collection, SCRAN's focus is on the use of that collection. In this respect, its activities approximate those of the cultural heritage and library institutions which are discussed above. Access to SCRAN's holdings and to associated support information and services will be available over the Internet via World Wide Web browsers.

*Rights management*

The funding agencies may manage rights as a means of further enhancing their influence over the life cycle of data resources produced by their grant holders. At NERC, the intellectual property vested in any data resources that are created by NERC employees are owned by the NERC itself, enabling the funding agency to determine their future disposition and use. Where data resources are created by NERC-funded third parties (e.g. university-based academic staff), intellectual property resides with the third party (e.g. the host University) but the NERC may attach as a condition of grant, a clause requiring that any such data are deposited with a designated NERC data centre and that that centre be given a non-exclusive license to distribute them for educational use. The NERC also takes pains through user licenses and other procedures to ensure that

appropriate educational re-use is made of NERC-funded data resources many of which have potential for commercial exploitation. Although mis-use is difficult to detect, the NERC carefully vets application for data access and is prepared to take action against users caught in transgression of the user license.

The SCRAN manages rights to similar effect. As a condition of grant, SCRAN acquires from its grant-holders a non-exclusive right to distribute any digital resources they produce, and to exploit those resources for commercial purposes. It also imposes user agreements on its subscribers to ensure that data distributed through SCRAN are used for educational purposes only.

### 5.4 The institutional archives

*Introduction to case study*

Institutional archives selectively build and manage unique electronic records which are generated by an organisation and retained by that organisation to document its activities. Institutional archives also make those records available as required by the record-generating organisation. Owing to the role that retained records play in documenting an organisation's activities, transactions, etc., and potentially as legal evidence, their authenticity is of paramount importance. Records which are modified in a manner which potentially alters their content or opens up new possibilities for interpretation threaten to misrepresent the organisation's activity.
Selection criteria and procedures are governed and enforced by the record-generating organisation which may also legislate how electronic records are created, documented, appraised for retention, transferred to the archive, and how and by whom they may subsequently be used. Given the archives' role in retaining materials which document an organisation's activities, its latitude is limited with regard to the changes which can be introduced into deposited data with regard to their format, contents or documentation. Such interventions threaten irreversibly to change or to bias the organisation's record. The institutional archive will have some influence over the organisation's acquisition policy, over its records management guide-lines, and over the appraisal process which it uses to determine what records to retain.

Interviews were conducted with representatives of two institutional archives, the Public Record Office (PRO) in the UK and the Center for Electronic Records (CER) of the National Archives and Records Administration of the United States (NARA). The focus on government archives may have biased the following assessment which, should be taken as partial until at least a selection of business archives are also consulted.

The NARA identifies, accessions, preserves and provides access to records of federal government which document the work of government and are of enduring interest. The CER is a department of the National Archive which deals with records in electronic for mat. It manages some 100,000 data files which together occupy more than 400GB of file store and the collection is growing all the time.

The Public Record Office (PRO) is the repository of the national archives for England, Wales and the United Kingdom. It has two programmes addressing the long-term preservation of electronic records in government in the UK, both in their early phases o f development: the Electronic Records in Office Systems (EROS) Programme which provides leadership across government in the UK in the management of electronic records in office systems; and the UK National Digital Archive of Datasets (NDAD) Programme which provides a framework for the preservation and use of government datasets in the UK. Both programmes are concerned with the archival preservation of public records of enduring interest. The former focuses on mechanisms for ensuring that documentary records are created in a manner which will facilitate their long-term preservation. The latter involves the preservation of very large structured datasets for which maintenance and user services have been contracted out to the ULCC as discussed above.

*Data creation and collection development*

<u>Data creation</u>

The institutional archives do not create data; they preserve data created and deposited with them by others. As such they have a substantial interest in influencing the behaviour of data creators since that behaviour will determine how, whether, an d at what cost any data deposited by them in the archive can be preserved in the longer term. In this respect, influencing data creators' behaviour is mission critical to the archive and is highly sought after through both formal and informal means. Formally, archives may influence data creators' behaviour through any influence they can exercise over the data generating organisation with regard to the development and implementation of records management policies and procedures imposed by that organisation over its data generating departments and members. Informally, archives may seek to influence data creators' by providing and promoting guide-lines to best practice in data creation, documentation, and management, and by offering an advisory function to data creators who seek to develop a data resource or to evaluate and prepare an extant data resource for possible deposit with the archive.

Both formal and informal influences are apparent at both the PRO and the NARA. Formally, both organisations exercise varying degrees of influence over the development and implementation of government policies with regard to the management of electronic records. Informally they prepare and disseminates extensive guide-lines regarding best practice and, as a matter of course, enter into dialogue with data creators as they seek to appraise and then prepare data for archival deposit.

Additionally the PRO and other national archives are beginning to focus their influence on software vendors, encouraging them to build record management functionality into software that is commonly used in government departments (e.g. electronic mail and document management software). This reflects recognition of the importance of the initial data creation process for long-term preservation, and the vital contribution that automated procedures can play in ensuring metadata is captured and retained with records

Data acquisition

Records held at the institutional archive are effectively those selected for retention by the organisation which it serves according to regulations established by that organisation. Such selectivity is as mission critical to an institutional archive as its influence over data creators' behaviour since it too will determine how, whether, and at what cost accessioned data can be preserved.

At the PRO, records which are retained for archival purposes are those covered by the Public Records Act. Guidance on the selection of records required for long-term preservation is given to record officers in departments and agencies in the Manual of Records Administration. Similarly, the NARA accessions records as required by statute and provides detailed implementation guide-lines to departmental officers. Although it is departmental record officers who are responsible for scheduling records for deposit, in the US, the NARA advised them, then reviews their schedules and either accepts the disposition of records as proposed or raising a counter position. Final decision about a record's disposition rests with the National Archive and with the CER where electronic records are concerned.

The advent of electronic information systems and the proliferation of electronic documents and other records has required extension and refinement of existing guide-lines prepared originally for paper-based materials. At the NARA revision has resulted in the development of a detailed policy and guide-lines regarding the appraisal, deposit, and management of electronic records. The PRO is developing similar policy and guidance. The criteria applied through the guide-lines in the appraisal of electronic re cords for long-term preservation are typically based upon a mix of content and technical criteria, and to some degree on an assessment of the costs involved in their long-term retention, particularly in the case of data from legacy systems. Content criteria may be equivalent to those used for paper-based ones and concentrate on records of enduring interest. Technical criteria reflect the archives' interest in preservation over the longer term and thus upon the use of software-independent data formats and media. Given the rigorous appraisal process, the archive's emphasis on records' long-term value, and the uniqueness of the archived records de-selection is not an issue. However it is recognised that electronic records are a

relatively new area. They have to be selected at a much earlier stage than traditional records, and therefore that review of electronic records may be more appropriate after a period of time has elapsed

*Data management*

Data structures

Here, too, the record-generating organisation's interest in long-term preservation, and in preserving the authenticity of its archived records are paramount in the requirements imposed upon depositors with regard to acceptable data formats and storage or data transfer devices which they must use to get material to the archive. Where electronic records from automated data processing systems are concerned, the overwhelming emphasis is on tried and tested software-independent data formats which also reflect the archives' focus on data resources which are one or more computer generations old. At the NARA, regulations require government departments to deposit such data as ASCII or EBCDIC files written onto 3480-class tape cartridges, 9-track tapes, or on CD-ROM. In a similar vein, typical accessions for the ULCC which acts as a data bank for the PRO typically include alphanumeric, field delimited ASCII data. Data deposited with ULCC for retention by the PRO are supplied as read-only unencrypted files which follow BSI PD 0008 code.

Electronic documents and other data types are more problematic. The National Archive permits the deposit of electronic documents as SGML-encoded ASCII files but has no requirements for digital images, digital sound, and other data types. The PRO, through its EROS programme assumes that all formats used for non tabular data (e.g. electronic documents, images, etc.) will eventually become obsolete, but also that some formats are likely to have a longer life than others. Such formats have been recognised in the EROS guide-lines as approved data preservation formats and include *de jure* and *de facto* standards such as ASCII, Postscript, and TIFF. By selecting such standards the PRO aims to reduce the frequency of data reformatting that may be required to migrate data through different standards and technology regimes, and thus to minimise the cost of data preservation.

Despite their reliance upon strict standard regimes, non-standard data may also be accessioned, for example, where they meet content requirements or where deposit is otherwise mandated. Data deposits at the ULCC on behalf of the PRO, for example, may include exceptional cases such as data stored in binary or proprietary formats (e.g. GIS data) or as alphanumeric data with binary formats embedded within them.

Where non-standard data are accessioned, every effort will be made to render them into some non-proprietary format with minimum content loss. Here, problems arise only data stored in proprietary or binary formats are insufficiently documented to make such transformation possible.

Data storage

Data storage will be contingent upon what access is required by the record generating organisation. Data requiring frequent access may be stored on-line or cached (a process approximating on-line storage for data which are actually stored off-line but frequently accessed and used). The ULCC has adopted a mixed scenario on behalf of the PRO. Its holdings exist within the public domain and accordingly will be accessible to members of the general public. Frequently accessed collections will be mounted on-line. Others will be distributed via ftp and on a range of suitable portable media. Both topics are dealt with in greater detail below under data use.

Data restructuring

The institutional archive's strict reliance upon robust software-independent data standards are intended to minimise the need to reformat data. Still, reformatting may take place under the following circumstances:

- when non-standard or process-dependent data are accessioned and are transformed into standard software-independent formats;
- when data stored according to previously approved formats (e.g. at the CER when older EBCDIC files stored on 7-track tapes) are migrated onto current ones (e.g. ASCII files stored on 3480-class tape cartridges);
- when data are reformatted to enhance or facilitated their mandated use (e.g. the ULCC adopts mixed distribution scenarios delivering data to end users on some portable media and making some data browsable via the Internet and the World Wide Web. Copies of data stored in software-independent standard formats may be taken and reformatted into non-standard or process dependent ones in order to facilitate their quick and effective Internet access).

Data documentation and description

Institutional archives tend to implement minimum level documentation as prescribed by the record-generating organisation. At the CER, the minimum level standards is imposed by regulation upon data depositors who, in many cases will also supply additional printed and electronic materials indicating how, why, and by whom a particular data resource or information system was created and used. Such documentation is verified by the CER upon accession - that is, the CER ensures that the minimum level documentation relates to the underlying data - using automated procedures which are also used to create a record for the title list. No further intervention is made by the archive with regard to data documentation owing to the volume of data accessioned and its concern over authenticity.

Similar procedures have been adopted by the PRO. A transfer form including an inventory of supplied information is provided for departments preparing materials for deposit. Here too, depositors, may provide other documentation in printed and electronic form. For data retained by ULCC on behalf of the PRO, the ULCC creates a standard catalogue record from the documentation supplied and enriches that record with information gathered by ULCC staff about the record's origin, structure, and uses. ULCC also compiles a brief history of the government departments and agencies depositing data, the first time such data are deposited by those departments or agencies. In all cases, these enrichments are intended to facilitate its secondary use.

Preservation

Institutional archives prefer migration as a preservation strategy. Given their reliance upon robust and stable software-independent data standards, migration principally entails copying data files periodically onto new storage media. Restructuring and reformatting is kept to a minimum as described above.

The potential value and importance of other preservation strategies are also recognised and currently under investigation at the PRO. Technology preservation, for example, is seen as a viable short-term option for the retention of computer hardware, though may have longer-term value for the retention of computer software. To investigate, the PRO is sponsoring a British Standards Institute (BSI) working group to look at extracting bundles of view, browse and extract functionality from any software and t o preserve that functionality with any data generated, stored by, or accessed from that software.

Preservation practices mirror those of the data banks (and, in the case of the PRO are fulfilled by a data bank for its large-scale datasets), though where storage media are concerned, extend to prescribed requirements for data transfer and data access , as well as for data preservation.

*Data use*

Use of the institutional archive's holdings is determined by the record-generating organisation and will be contingent upon that organisation's needs. In all cases, the depositing organisation will require access to its holdings for pre-determined periods of time. In the case of the national institutional archives accessioning records of government, data are in the public domain (though

in the case of the PRO under Crown copyright) and accordingly must be made accessible to end users, though with some restrictions (for example, the 30-year rule which applies in the UK), and confidentiality issues. What that access will entail will be determined in part by the record-generating organisation's requirements, the nature of the data, and the funding and technologies available to the institutional archive.

Information about data held by ULCC on behalf of the PRO, and about those available from the CER are or will be disclosed through on-line catalogues and title lists, respectively. In the case of the CER the title list holds information about some 15% (15,000) of the deposited data resources and is available on-line, in printed form, and via ftp as a text file. Other references to the CER's holdings exist formally and informally in various references and reports which may be available to users either on line or in print. Information about electronic documents in the PRO's EROS programme are accessible via workstations in the PRO's reading room in which they are represented with the same information that is supplied for paper records, and there are plans for the development of special on-line finding aids.

With regard to data access, the CER makes copies of electronic records files available on a cost-recovery basis, on researcher's medium of choice, and written according to the researcher's technical specifications. In addition, staff of the CER will in certain cases progress user-specified queries against data resources where those queries are viable and based upon the documentation supplied for that data resource. The CER will also, when feasible, conduct a user specified query where that query is based upon the documentation supplied for a file. The PRO has adopted a mixed strategy. Some of its holdings will be made accessible via the Internet and standard Web browsers, and all of them not otherwise restricted will be available to users on portable media and in file formats to be determined.

The user support supplied by the institutional archive will reflect the nature, extent, and information requirements of its depositing and user communities. At the CER and the PRO, both depositing and using communities are broad (the former extends to all agencies of national government and the latter to a theoretically unlimited user population) and range widely in their understanding of the archive's functions, procedures, and holdings. Depositing and using communities will also vary in their sophistication with regard to computer technologies and methods. Accordingly, both agencies supply richly detailed information via electronic and other means. That information tends to target naïve users rather than sophisticated ones, though there are short-cuts through it intended for the latter.

For depositors, the support materials tend to be oriented around records management regulations and provide guidance to data creators in their implementation and use. User documentation describes the archives functions and its holdings and guides individuals in the acquisition of data resources (or in the formulation and presentation of queries to archive staff which are based on those holdings).

*Rights management*

Rights management will be contingent upon the terms and conditions of use laid down by the record-generating organisation and any national legislation that applies. For the CER, rights management is not an issue since records of federal government are in the public domain. Records which are restricted by the federal government (or portions of data resources which are restricted) are not released to users. PRO records are covered by the Public Records Act and by Crown copyright and standard Crown licenses from Her Majesty's Stationary Office (HMSO) are used.

### 5.5 The "academic" data archives

*Introduction to case study*

Academic data archives selectively develop, maintain, and encourage re-use of unique data resources which are of interest to particular end-using communities and which are drawn from a wide variety of depositors. These resources have typically been electronic texts or datasets but

are increasingly extending to other data types (e.g. digital images, sound, film, and GIS). Like the digitisers, the academic data archives are influenced by the information needs of the end user communities that they serve. Unlike the digitisers, the academic data archives tend on the whole to serve relatively small or at least well defined end-user communities whose members have a relatively deep and sophisticated understanding of the data and their uses. Another difference is that academic data archives exert less influence than the digitisers over how data are created. In this respect, they are closer to the institutional archives and, like the institutional archives, will attempt to influence data creators in both formal and informal ways. Unlike the institutional archives, the academic data archives rely more for their collections on voluntary and/or non-exclusive rather than on statutory deposit, - a feature which tends to minimise the direct influence they may have over depositors behaviour as data creators. Accordingly, academic data archives will typically devote resources to enriching deposited data to ensure they are maximally useful to and usable by end users.

Interviews and case studies have been drawn from the Data Archive (DA) and the Arts and Humanities Data Service (AHDS), established for the social sciences and the arts and humanities, respectively. They have also been informed by further interviews with the Space Data Centre of the Rutherford Appleton Laboratory and the NERC as examples of academic data archives established for the scientists. The DA was established at the University of Essex in the 1960s by the Social Science Research Council (SSRC, now the Economic and Social Research Council, ESRC) to acquire, preserve, and disseminate social science survey data, particularly that resulting from research funded by the SSRC. Since the scope of its collections development aims have been extended considerably to include a broader range of data of interest to social scientists and, since the early 1990s of interest to historians and psychologists as created by SSRC/ESRC grantholders, UK government departments, or others. The AHDS was founded in 1995 as a distributed organisation managed as a single entity and comprising five subject based Service Providers in archaeology, history, the performing arts, the visual arts, and textual studies. Each of the Service Providers collects, manages, and encourages re-use of digital resources which result from or support research and teaching in their particular subject areas. Just as the Data Archive works closely with a funding agency which invests in work of interest to the social sciences, so does the AHDS work with funding agencies and charitable bodies involved with the arts and humanities. Currently, the Carnegie Trust for the Universities of Scotland, English Heritage, the Humanities Research Board of the British Academy, the Leverhulme Trust, and the Wellcome Trust's History of Medicine Programme either recommend or require their grant-holders to offer data produced in the course of their research to the AHDS for deposit. The AHDS's Archaeology Data Service is also recognised by the NERC as a data centre for science based archaeology data.

*Data creation and collection development*

Data creation

Academic data archives are as eager as institutional archives to influence the behaviour of data creators and for the same reason. How data are created will determine how, whether, and at what cost they can be preserved. The academic data archive's influence over data creators, however, is less than that exercised by the institutional archive. Academic data archives rely largely for their collections upon voluntary deposit by data creators who may not be subject to any standards regime other than research criteria,. Nor are data creators ideally positioned to implement such standards retrospectively since they typically offer their data to an academic data archive after their active interest in those data has ceased.

The academic data archives' influence over data creators is therefore largely informal. It tends to be exercised crucially through their relationships with funding agencies (which can recommend the standards and best practices identified by the academic archives) or through a range of informational and instructional activities. The latter will typically seek to raise awareness about the long-term research and educational value of data resources created in the course of research, and about the importance of standards best practice in data creation in terms which appeal to the data creators' self interest. Thus, the adoption of data standards, for example, is promoted by the AHDS in terms of their benefit to the depositor. Such standards, AHDS guide-line s suggest, ensure that data can be meaningfully accessed despite changing technical regimes. As such, they

support secondary use from which, the full scholarly investment in a data resource can be realised.

Data acquisition

Academic data archives are highly selective in their accessions and use both content and technical criteria to maximise their holdings' re-use value to a particular end-user community. Content criteria are shaped by the archives':

- formal and informal assessment of what data resources end users require and seek actively to accession these where they exist;
- formal and informal assessment of research and teaching trends - academic data archives will attempt to predict the emergence of such trends and develop collections of data which can support them;
- interest in developing special collections - such collections may fill gaps in the archives holdings or supplement existing holdings to create a critical mass of material focusing on particular topics or themes. The Data Archive has developed exceptionally rich data series including annual or periodic Family Expenditure, General Household, and National Food Surveys, extending back into the 1960s, and has opinion poll data extending back to the 1930s. The History Data Service (an AHDS Service Provider located at the Data Archive) has extensive nineteenth-century British census and census-like data. The development of special collections may entail more active recruitment of relevant data resources, or, funding permitting, the development or enrichment of such data resources by the archive.

Technical criteria help to determine whether and how and at what cost a data resource may be used by others and preserved by the archive over the longer term. In general, the academic data archive will give preference to data resources which conform to a suite of specified technical and documentation standards chosen to reflect the kind of data being accessioned (electronic texts are formatted and documented differently than digital images), and the information requirements of end-using communities. Significantly, academic data archives will not necessarily reject data resources which promise long-term secondary use value but fail to meet preferred technical criteria. Rather, the archive will define a strategy whereby such a resource may be accessioned and, funding permitting, appropriately enriched, transformed, or improved.

The academic data archive's interest in secondary analysis for research or re-use in teaching also influences the accession process. Documentation supplied with deposited data will be verified against the data themselves and, where deemed appropriate, improved or enriched to standard in order to facilitate its secondary use. The Oxford Text Archive, for example, will ensure that electronic texts within its collections are accompanied by an SGML header conforming to the Text Encoding Initiative's specifications and to in-house cataloguing practices. The Data Archive similarly will prepare comparable catalogue records for its data resources, and ensure that data resources are accompanied with user-supplied codebooks and other documentation. In some case s, the Data Archive will also prepare user packs offering guidance in the use of particularly complex data resources which are in high demand.

Similarly the academic data archive may intervene in the data's structure. The Oxford Text Archive prefers the use of TEI-conformant SGML and will undertake to introduce this encoding where possible for resources in its collections. The Data Archive principally accessions alphanumeric data in a range of tabular or list-like formats but prefers to store data in the SPSS file format owing to the ease with which the SPSS software may be used to produce subsets and to write data into a variety of different formats as may be required for delivery to end users. Where possible it will transform data resources into the SPSS file format upon accession.

De-selection is not usually an issue for the academic data archive although version control is, particularly where the archive holds data resources which are either periodically amended or updated by their creators or subject to further development in the course of their secondary use. Where deposited resources are superseded by newer versions, older versions will be maintained though their distribution may be restricted. It is important to note that version control adds a level of administrative complexity

*Data management*

Data management in the academic data archive reflects its dual emphasis on secondary use and long-term storage, and are contingent upon what kinds of data an archive caters to and the end-user scenarios it wishes to support.

Data structures.

The academic data archive is likely to hold copies of its holdings in different formats and reformat accessioned data accordingly. Such formats may include:

- that in which the data are deposited;
- that considered most appropriate for the data's archival management;
- that considered most suitable for distribution to end users.

Inevitably, reformatting adds a cost to the academic data archive which will be considered when a data resource is evaluated for accession. That cost can be reduced substantially where data creators can be encouraged to deposit data in approved formats , though there are severe limits on the extent to which the adoption of such formats can be encouraged.

The Data Archive's collections currently consist primarily of alphanumeric data stored in tabular or list-like formats and delivered to end-users on a range of portable magnetic media and by ftp, and in a range of different formats. The Archive accordingly accessions appropriately documented data in any number of formats (e.g. as delimited ASCII files, as internal format files from known database software). It prefers SPSS file formats as a management format, however, owing to its flexibility, and may convert accessioned data into that format. Additional copies of the data in different formats will be created in response to user demand. Thus, a data resource deposited as a SAS export file may be stored as a SAS export file (the accession format) and co pied at accession into an SPSS export file format (the management format). Where a user requests the data resource as a SIR file, it may be reformatted as required for distribution and a copy of the SIR file may be stored against future requests for the data in that format.

The Oxford Text Archive accessions electronic texts as ASCII files employing a range of different mark-up (encoding) schemes and conventions. Preferring TEI-conformant SGML, it will where possible impose a light SGML encoding onto all deposits, storing the original file as deposited, and the SGML-encoded version. Texts are distributed as SGML-encoded files but may in some circumstances be transformed or re-encoded as appropriate to their use with particular application software (e.g. HTML-aware browsers).

Data storage

As with the institutional archives data storage scenarios reflect the data distribution ones that the academic data archives wishes to support. On-line storage is paramount for Internet access, but also for frequently requested resources which are distributed on portable media.

Data documentation and description

Given its orientation toward secondary analysis, the academic data archive stresses the use of richly descriptive documentation which may be considered at several levels:

- a brief catalogue record with which users will be able to locate a resource and to decide whether it warrants further investigation;
- more richly descriptive information about the resource's contents, context, creation, provenance, and structure, and about the terms and conditions of its use which may enable the user to determine whether or not the resource is worth acquiring for a particular research or teaching purpose;
- supplementary information as and where required to enable the user to mount and analyse or use the resource for their own purposes (e.g. code books for variables used in a database, information about any sampling strategies used to create a resource, etc.).
- In some cases, for example, with an SGML-encoded electronic text, all of the above

information is contained within a single electronic record or record component (in this case, within the SGML header). In others, it may be available in electronic and printed documents (for example, with a social science database for which a catalogue record exists independently of a detailed study description and codebook) which may be stored independently of the data themselves.

Depositors are best suited to the supply of required documentation. Consequently, the academic data archive will encourage them to document their data by providing detailed guide-lines and professional assistance. Despite this pro-active approach, experience suggests once more the limits to depositors' compliance with archive approved standards. Preparing documentation which is appropriate and sufficiently detailed for secondary analysis and for preservation, accordingly tends to involve both depositor s and archive staff and frequently has substantial resource implications for the academic data archive. These resource implications are typically considered by the academic data archive when evaluating a data resource for possible accession.

Data preservation

Academic data archives demonstrate a preference for data migration as a preservation strategy; a fact reflected in the technical criteria and preferred data standards against which potential deposits are evaluated and which depositors are encourage d to adopt.

Preservation procedures reflect the infrastructure and technologies available to the archive and vary substantially. In the best case, preservation procedures reflect those used by the data banks and the institutional archives. In the worst case, they rely upon more minimalist practices which extend no further than regular back-ups of data stored on-line.

Remote management

Academic data archives' collections may have physical and virtual holdings. Physical holdings are those for which the archive is the principal managing service. Virtual components are those which are available from the archive but which are managed off site. Virtual holdings may be mirrored at the archive but have a primary archival home at another organisation. The Data Archive, for example, distributes data resources which are held by the Inter-University Consortium for Political and Social Research (ICPSR) in the US. The first time an ICPSR data resource is distributed by the Data Archive, a copy of that resource may be kept by the archive against subsequent user demand. Despite its holding a mirror copy, the ICPSR retains primary curatorial responsibility over that resource. Virtual holdings may also be managed entirely off-site though references to them made available to them through the academic data archive's catalogue. The Archaeology Data Service, for example, includes in its catalogue references to data resources which are maintained off-site by the Royal Commission for the Ancient and Historical Monuments of Scotland.

Typically, remote data management relies upon agreements between the academic data archive and third-party data managers. Those agreements will seek to assure the academic data archive that remotely managed data conform to minimum level content and technical requirements as normally applied by the archive in selection of its own physical holdings. They will also seek assurance that the data are maintained and distributed in a manner which conforms to and fulfills the archives' own data management and us e strategies.

*Data use*

Academic data archives typically make information about their holdings available through [on-line] catalogues, and may adopt one or several of the following data distribution scenarios:

- data resources may be accessed and manipulated over the Internet via Telnet sessions or World Wide Web browsers;
- data resources may be distributed to end users over the Internet via ftp (File Transfer Protocol);
- data resources may be distributed to end users on a predefined range of portable magnetic media.

However they are made available, the re-use of data resources within the academic data archive will be contingent upon a number of factors including their relevance (content), quality, structure, and documentation. As we have seen, the academic data archive strives to enhance the usefulness of its holdings by being highly selective about what resources it acquires and by enriching deposited resources (e.g. by enhancing their documentation, or by reformatting their structure). The academic data archive may also seek to improve and enhance the usefulness of its holdings by influencing the behaviour of individuals and organisations which are responsible for the creation of potentially depositable resources, raising awareness about the benefits to be had from secondary use and of the crucial importance of appropriate standards and best practices in the creation of re-usable data resources. The academic data archive may play a similarly active role amongst potential user communities, raising awareness about the benefits which accrue from secondary analysis, and providing guidance to best practice in its conduct.

With regard to support, the academic data archives, like the institutional ones, focus their attention on their two principal (and potentially overlapping) user communities including depositors and users, respectively. Both are supported with documentation which may be available on-line and/or in printed form. Depositors are typically supplied with detailed information about the deposit process (including detailed information about the evaluation criteria employed by the archive and the data formats an d documentation standards that it either prefers or requires), and about the terms and conditions under which data are deposited (see below). Depositors are also encouraged to access and adopt those technical and documentation standards which are promoted by the archive in order to support its dual aims of preservation and use.

Users will, at a minimum, have access to information about the archive, its holdings, and about procedures for and terms and conditions applying to access to the archive's holdings.

*Rights management*

The academic data archive acts as an intermediary between depositors and users, and will use inter-related depositors licences and user agreements to ensure that any restrictive terms and conditions imposed by depositors upon are adhered to by *bona fide* users.

Given their emphasis on secondary use, there will be a strong emphasis in the licensing on protecting the moral rights, academic integrity regulation citation of the resource, and the impartiality of data creations and funding agencies, than may occur elsewhere. For resources derived from censuses, social surveys, or other government sources, measures to protect the confidentiality of individuals and institutions will also be required. Academic data archives may also be reluctant to accession resources whose re-use at least for personal and educational purposes, is severely restricted, and to distribute resources to individuals and institutions which cannot prove their *bona fides* as educational users. Their pre-eminent reliance upon voluntary deposit, however, may have an opposite effect, encouraging academic data archives to restrict or at least guard access very closely using a range of authentication and other mechanisms, so as not to deter depositors from offering their data in the first place.

### 5.6 Legal deposit or copyright libraries

*Introduction to case study*

Copyright libraries have a statutory obligation to maintain and provide access to certain classes of non-unique information objects whose deposit is legally prescribed and enforced upon the objects' producers. Copyright libraries may supplement these core holdings through voluntary deposit and through acquisition (either through subscription or purchase), both to support their national roles in preservation and to fulfill other functions such as the provision of information both on site and through inter-library loan. Their use of the framework is governed by mandated deposit, the need to influence how publishers create electronic information objects, their orientation toward long-term retention and preservation of information which are covered by legal deposit materials and toward permitting the (possibly limited) secondary use of those

materials (Mackenzie Owen 1996, Mandel 1996).

It is important to note, however that legal requirements to deposit electronic publications with copyright libraries is relatively new and to date, only a small number of countries have appropriate legislation in place. Interviews for this study were conducted with representatives of two copyright libraries, the British Library (BL) and the National Library of Australia (NLA). Both reside in countries where national legal deposit legislation has not yet been amended so as to extend to electronic publications but where such amendments are expected in future (in Australia, several State Libraries are already entitled under State legislation to receive electronic publications). In preparation for the amendment of legal deposit legislation, both the BL and the NLA have conducted research into the accession, preservation, and use of digital information objects, and at the NLA to the establishment of a pilot data archive.

The discussion which follows offers a general examination of the framework from the perspective of the copyright library, identifying issues of distinctive concern, and elaborating upon them with reference to digital preservation projects launched by t he libraries with CD-ROMs (the British Library) and with on-line publications (the National Library of Australia), respectively.

The BL's Legal Deposit CD-ROM Demonstrator Project was established to assess existing methods of registration, record creation, and bibliographic service provision for one class of electronic publication (CD-ROMs) likely to be amongst the first categories of non-print publications to be deposited with the library if legal deposit legislation is extended to electronic publications. The Demonstrator also undertook preliminary investigations into requirements for preservation of and access to deposited CD -ROMs.

The NLA's PANDORA project was established to capture, archive, catalogue, and provide long-term access to significant Australian on-line publications which are selected for national preservation. It is conducted through the NLA's Electronic Unit which scans the Internet for online publications of national significance and selects for archiving those which meet criteria set out in guidelines prepared by the Selection Committee for Online Australian Publications (the SCOAP guidelines).

*Data creation and collection development*

Data creation

Copyright libraries are expected to retain publications which are deposited under the terms of legal deposit legislation for the dual purposes of preservation and access. As such, they must be prepared to accession electronic publications combining different forms of data such as images and text and which can be hyper-linked to other resources. They are deposited in a wide variety of different formats, on a similarly wide range of media, and which are provided with varying degrees of documentation. In this respect, copyright libraries share certain characteristics with the data banks. Unlike the data banks which preserve data at the bit-stream level and leave content considerations to the depositor, the copyright library must also confront issues pertaining to the preservation and re-use of a data resource's content, appearance, and functionality. That is, it is or will be required to ensure that the content, appearance, and functionality of deposited electronic publications can be meaningfully rep resented on and accessed from contemporary hardware and software platforms. In this respect, the copyright library shares characteristics in common with the institutional and academic data archives. Unlike the archives which ensure that the content of its data holdings remain viable by exercising some direct influence over how those holdings are created or by restructuring or enriching those holdings at accession, the copyright library exerts no such influence and the format and type of data deposited may mitigate against restructuring at accession. Additional constraints may be imposed upon the copyright library by the terms and conditions of legal deposit. Where those conditions permits only single user access to deposited holdings, for example, certain strategies for data management and access which may be preferable owing to their efficiency and cost effectiveness (those relying on centralised storage of network accessible information) may be prohibited.

Of course, the difficulties confronted by the copyright library are less severe where electronic publications are acquired through voluntary deposit or through purchase or subscription. There the copyright library may develop and employ selection criteria to ensure that only those materials are accessioned which can be managed effectively.

Accordingly, the copyright library is or will be compelled to implement the widest possible range of data preservation and access strategies appropriate to the diverse preservation and access requirements imposed by the diverse characteristics of its electronic publications.

Given their lack of any formal influence over electronic publishers whose products are or are likely to be covered by legal deposit legislation, copyright libraries will typically:

- demonstrate an interest in reserving the right to select from amongst deposited electronic publications those which they will retain for the purposes of preservation and/or access;
- demonstrate a significant interest in working together with publishers to identify and promote those standards and best practice which help to ensure that electronic publications can be migrated across platforms and between individuals at minimal cost and with minimal loss of information content;
- demonstrate a significant interest in working together with government to ensure that the preservation and access obligations that are or may be imposed upon the library by legislation requiring the deposit of electronic publications are (a) achievable and (b) adequately resourced.

Data acquisition

The selection process is likely to involve both content and technical criteria. Content criteria may reflect those used in print materials and differ depending upon whether accessions are to be acquired by voluntary deposit, by purchase/subscription, or by mandatory deposit (where selectivity of such materials is permitted). Additional technical criteria may also be applied to electronic publication as a means of assessing the extent to which they may be retained and made accessible to readers in light of any technical and financial constraints. A key problem encountered by both the BL and the NLA was that of identifying and then "claiming" electronic publications for deposit. Mechanisms used to identify and claim print publications (comprehensive lists based on standardised and unique publication identifiers, regularised information exchanges with appropriate copyright and other agencies, publishers awareness of the copyright library's mission with regard to deposit) were either non-existent o r poorly developed for electronic ones. Where on-line publications were concerned, Internet search engines offered only moderate assistance in the identification of appropriate electronic materials.

"De-selection" is likely to reflect library retention policies as they apply to particular collection levels and classes of objects within its collection. The British Library's retention policy for example, recognises three categories of object: heritage and unique; long-term research; and current (including duplicates etc.).

The on-line publications in the PANDORA archive are equivalent to the heritage and unique level at the British Library and reflect the NLA's underlying philosophy for all publications - to ensure comprehensive collections of Australian materials are preserved in perpetuity. With regard to digital on-line publications, the NLA looks towards a co-operative national model where the State and possibly other libraries (e.g. large university libraries) and organisations will share responsibility within a national framework for archiving and providing access to electronic publications. The SCOAP guidelines make a contribution toward the development of that framework (as, indeed, does the PANDORA project which acts in some respects as a test case) and provide selection criteria. Generally, content criteria prevail and there is no discrimination against on-line publications on the grounds of their technical formats (though see below) or of the costs involved in accessioning and preserving such materials.

Implementing the guide-lines proved to be more resource intensive than originally expected. Identifying on-line publications was an initial hurdle, and only moderate help was available from Internet search engines and logically structured gateways. The evaluation of publications once

discovered was also resource intensive, requiring detailed evaluation of their content and technical characteristics. Given the inherent diversity of electronic publications, no single evaluative method could be employed - the information required to support it could be available (if at all) in different places and accessible via different means with different publications.

Once a publication is selected for inclusion in the archive, a preservation strategy is devised for it. Costs and file formats are not at issue though with publications involving idiosyncratic formats or a publishers' proprietary (non-third-party) search/retrieve/access software, PANDORA may respectively consider migrating the data into more standard formats and acquiring preservation copies of the proprietary software used for its delivery. The Pandora project does not intend de-selection; a stance which reflects the selectivity which guides decisions to archive on-line publications and because of its long-term and national perspective.

The BL's CD-ROM Demonstrator Project was based upon a sample of 133 titles selected to represent a range of formats and subjects. Through various tests, a number of selection and accessioning issues distinctive to CD publications were identified including:

- Resource identification. Extant lists of CD-ROM publications are not comprehensive, unique identifiers are not used universally with CD publications as they are with printed ones; mechanisms for exchanging information about CD publications with appropriate copyright agencies are less developed.
- Resource selection. CD-ROM publications are often produced in different versions, for example, for use with different operating systems. Guidelines were required to govern selection of such CDs.
- Evaluation and accession. Information required to evaluate a CD-ROM based publication was not comprehensively available from the printed material supplied with it. Accordingly CDs had to be loaded and "read", a requirement which forced to the fore equipment and technical support issues which were simply not present with regard to the evaluation and accession of printed publications. CD-publications also require a number of evaluation and accession procedures for which there are no direct equivalents with print publications. These included virus checking and crucially "preservation checking". The latter is crucial. A printed publication is typically placed in storage after accession; its preservation requirements are then determined at a later stage. With electronic publications, preservation requirements need to be assessed at accession as these will determine or at least influence storage and access decisions. With CD publications, technical details are required about:
- the disc's chemical composition (determines rate of deterioration);
- operating system requirements;
- application software;
- file structure;
- data structure or architecture.

Research conducted for the CD demonstrator project demonstrated that operating system requirements were available by reading the packaging materials and that software requirements and file structures could be obtained by loading and reading the CD. Information about the disc's chemical composition and its data structure or architecture were not available and not necessarily obtainable from the publisher, particularly those who sub-contracted technical development work.

*Data management*

Data storage and data structures

In the copyright library, data storage and data structuring reflects preservation and access strategies. With PANDORA, data are structured as they are accessioned. In order that their look, feel, and content may be preserved. The fact that PANDORA is dealing with web-based publications whose construction and presentation are constrained by the web and by the formats and structures appropriate to it is a further deterrent. In most cases, data prepared for web-based delivery are prepared according to *de jure* or *de facto* standards which are presently widely recognised by appropriate software.

Data may be restructured under the following circumstances. Where published data are compressed, PANDORA will normally ask for utilities used by the publishers to uncompress them so they may be stored in the archive in an uncompressed format. Where published materials are "locked", PANDORA will normally ask that publishers provide that archive versions be "unlocked" versions for the archive. Where published data exist in idiosyncratic formats or are presented in ways which would make archived versions difficult to access or use, PANDORA may consider restructuring them. Note that cost considerations involved in accessioning and archiving a data resource would neither force nor inhibit data restructuring.

With CDs' the difficulties involved in determining their architecture makes emulation the most appropriate long-term preservation strategy. Migration is likely to be considered as a potentially less costly option for those publications whose architecture is known and relatively simple. In either case, CD publications would have to be restructured periodically to reflect changing technical regimes; comprehensively so where emulation is concerned. In addition, all CD publications will need to be copied from time to time onto newer CD platforms given their (often indeterminate) volatility.

Use considerations also apply, particularly where legal deposit legislation imposed the same single-user access constraints on electronic publications that are imposed upon printed ones. In such cases, readers may require physical access to the CD for use in a reader-accessible computer. Consequently, the physical location of a CD may become a storage issue as frequently used CDs would arguably need to be housed in proximity to an appropriately equipped reading room.

Data documentation and description

Cataloguing standards used by the copyright library for electronic publications are likely to reflect those used with printed ones, and thus to rely upon MARC and accompanying standards (e.g. in the English speaking world at any rate, on Anglo-American Cataloguing Rules, Library of Congress Subject Headings, and on the Dewey Decimal Classification system). The approach promises to integrate information about relatively new digital information objects with that already available for printed ones, an d thus to fulfill a national bibliographic function which is typically performed by the copyright library. The adoption of MARC has also been validated by research, successful thus far, into the use of MARC extensions to document distinctive characteristic s of digital information objects. Cataloguing practices are also likely to continue little changed with digital objects, with records generated by in-house staff, though some may be requested from publishers by cataloguing staff, notably about an electronic publication's technical characteristics, and, where on-line publications are concerned about patterns of publication, terms and condition of use, etc.

PANDORA prepares a full MARC record for archived publications and enters it onto the National Bibliographic Database. A catalogue record is also entered into the NLA's web-accessible OPAC through which users have integrated access to the NLA's diverse and mixed media holdings. In addition, a title entry pages are manually created for archived publications providing information about the publication, including a copyright statement, terms and conditions of use, and, where serials are concerned, information about the number of issues and their pattern of publication. Critically, the title entry page also provides a link to the publishers' home page encouraging users to use the current rather than the archived version.

In its CD-ROM demonstrator project, the BL found that UKMARC, AACR2, Library of Congress Subject headings and Dewey Decimal classification were appropriate and that records could be generated for CD-ROM's in about twice the time required for printed publications, principally owing to the need to read the CDs (and thus to load and then de-install appropriate software).

Data preservation

Preservation strategies appropriate for electronic publications are substantially different than those which are appropriate for books, not least of all because such strategies need to be worked

out when (even before) an electronic publication is accessioned on the basis of the publication's technical characteristics which are not always transparent or even available. Inevitably, the copyright library will have to adopt a range of different strategies as appropriate to the diverse kinds of electronic publications in their collections, and to the terms and conditions imposed on the library with regard to their maintenance and use. Migration may be preferred as a least costly option with data resources for which it is appropriate. Emulation will also be required particularly with electronic publications produced in non-standard and undocumented proprietary formats and/or those implementing proprietary software or hardware. There are a number of outstanding concerns, however:

- about the prospects for preserving some electronic publications, for example those fitted with hardware or software keys which time out and render the electronic publication inaccessible after a given period of time;
- about the physical security of items such as CD-ROMs which need to be handled physically by readers (e.g. those whose content cannot be migrated onto a central computing server for network distribution), and which, as such might be removed from the library or damaged through use;
- about the absence of clear cost models for any of the preservation strategies (but particularly for emulation) that the copyright library is likely to be required to adopt given the extensively diverse electronic publications that will be deposited;
- about the relative shortage of information about financial resources likely to be available in the longer term as necessary to determining appropriate and achievable preservation strategies for electronic publications accessioned now.

With regard to preservation procedures, copyright libraries may currently lack the large-scale infrastructure needed for mass computer storage and access available at the data banks and the institutional archives reflecting their historic orientation t o paper-based materials and other artefacts. The development of that infrastructure or of partnerships with agencies which maintain it may become an issue for copyright libraries in future.

PANDORA has adopted both short-term and long-term preservation strategies which are worked out for electronic publications based upon the technical evaluation which takes place before accession. Short term strategies involve routine and cumulative back -up of the on-line filestore to DLT cartridges. Long term strategies involves data migration through changing technological regimes. There are also potential roles for technology preservation and for emulation where on-line publications are only accessible via proprietary (non-third-party) software developed for and used by the publisher. In these cases, PANDORA would seek to work with the publisher to ensure that the software is migrated across changing technological regimes or that the look, feel, and content of the publication can otherwise be emulated. (Note, PANDORA does not yet have direct experience of emulation).

The preservation of CD-ROMs has already been addressed above. Migration may be appropriate for CDs with relatively simple data architectures. Most, however, are likely to require emulation. Given the volatile nature of many CD-ROMs as data storage media over the longer-term, all CD publications will need periodically to be copied onto new discs.

Remote management

Remote management may become an issue for copyright libraries, particularly where electronic publications covered by legal deposit legislation are dynamic, that is, available from a third-party supplier in a form which is constantly being amended and updated. The serial publications archived by the PANDORA project are examples of dynamic publications. Others include on-line reference materials such as bibliographic databases, which are periodically updated with new entries. Another example is the national large-scale mapping of the UK held in the Ordnance Survey's National Topographic Database. The Ordnance Survey is the UK's national mapping agency and previously produced traditional printed editions of large-scale maps. With computerisation, maps are now produced on demand for customers and the database is constantly updated with new survey information. The copyright libraries in the UK are currently discussing with the Ordnance Survey how regular snapshots of the national map-base could be archived for long-term preservation and access.

There are at least two strategies for dealing with such data as are covered by legal deposit legislation. One, adopted by PANDORA with its serial publications, is to create an archive copy of new issues as they are added to the serial. If applied to dynamic databases, the same strategy would entail the periodic creation of snapshots of the data resource for archival storage.

A second strategy is to manage the data remotely according to agreements struck between the copyright library and the third-party data supplier. In such cases, the academic data archive's approach to remote data management may be instructive.

*Data use*

Data use in the copyright library is likely to reflect and to integrate with practices pertaining to printed publications. It will also be contingent upon terms and conditions of use imposed by depositors or, where publications are deposited under copyright legislation by the terms worked out between the libraries, government, and publishers. Supporting access to electronic publications also make a range of new demands on the copyright libraries. Reading rooms need to be equipped with appropriately enabled workstations and the workstations need to be supported technically. Users will require adequate documentation and support services and possibly even training given the broad public which the library generally serves.

The PANDORA project makes information about its archived holdings accessible from the OPAC which supplies a hypertext link from the OPAC record to archived holdings' title entry pages which themselves encourage users to link to the publishers' site where this is still available on line. The holdings are made available via the Internet and web browsers, either from the publishers' site (the preferred port of entry for users) or the archive mirror of them.. Currently, all of the publications in the PANDORA archive are freely available on the Internet, and are also freely available items in the archive. One of the PANDORA business principles relating to use is that the archive will be available free of charge to users. However, once the archive contains titles on which the publisher imposes a fee for access, it is recognised that some restrictions on access will need to apply in order to avoid undermining the publisher's commercial interests.

Users are supported with information on the title entry pages and with on-line information made available by PANDORA about the archive and its use. Information about the archive and its activities is currently supplied to potential depositors individually by email, though formal depositors' guidelines or information packs may be considered.

With CD-ROMs, usage will be dependent upon what a library will be permitted to do by national legislation or rights holders in terms of providing networked or single access as with printed books. Single use can introduce administrative problems for monitoring users, usage, and ensuring non-disappearance of or damage to portable format e publications. There are also implications for reader services in terms of staff, equipment, training, technical support, user support, storage, and retrieval for users.

*Rights management*

For the copyright library, rights management practices are likely to be prescribed by any copyright legislation and will reflect the extensive discussions which will take place between government, publishers, and the copyright libraries that inform that legislation. In general, rights management practices will need to mediate between publishers who may seek to restrict access to deposited materials, users who may seek to extend access to published materials, and the copyright library which may seek some influence over the construction and management of copyright materials in order to ensure that they can be preserved over the longer term and made accessible to end users under what ever terms and conditions ultimately apply.

Currently the BL's practices with its CD-ROM collections reflect those in place for its print collections. That is, unless otherwise negotiated by the BL, CD-ROMs are not available to users for concurrent use or for remote access.

At the NLA, the title entry pages provided for each publication in the PANDORA archive provides a copyright statement drawing the users' attention to their responsibilities, and offering a link to the publishers' copyright statement in the most recent version of the publication in the PANDORA archive. PANDORA is also developing a voluntary deposit deed for use with publishers who wish to impose some restrictions on access to archived copies of their publications; the deed to be based upon one which is currently used for physical format publications.

## 6. Implementing the framework. A guide to best practice

These summary recommendations outline best practices to be adopted in implementing the framework with a view to the creation of data resources which may suit the purposes for which they are intended, and/or whose content, appearance, and functionality may be cost effectively preserved over the longer term. Only a sub-set of these recommendations may be relevant to any single data creation, collection or preservation initiative, depending on its role and its intended aims. The recommendations are therefore generic ones and it is intended they should be used with the case studies and bibliography, resources and references, which provide pointers to detailed guidance and application-specific practices.

1. Data and collection design. Whether data resources are being created, accessioned physically into a collection or rendered accessible to a defined user community, rigorous evaluation of the intended resources is essential to ensure that they serve the purpose for which they are intended at an affordable cost. Data creators, digital collection developers, and data archivists should be advised of best practice in the evaluation of potential data resources. That evaluation will at a minimum entail examination of the resource's:
   1.1. Contents, scope, relevance to a defined user community or purpose
   1.2. Technical characteristics (structure) and how these bear the resource's fitness for use by an intended community or other purpose, and upon its long-term maintenance and preservation
   1.3. Documentation (whether this is sufficient for its intended use and long-term maintenance)
   1.4. Legal terms and conditions which attach to its management and use

   Such an examination will take place in light of the evaluating organisation's mission, and the funding and technologies which are available to it.

2. Data creation.
   2.1. The principal influences over how and at what cost a data resource may be used, and over how, at what cost, and even whether it may be preserved are exerted when the data resource is created. Data creators need to be aware of this influence and accordingly of their role in determining the direction of a data resource's future life course.
   2.2. Data creators who are interested in ensuring the secondary use and preservation of their resources should be advised to adopt standards and best practices. Such standards and best practices reduce the cost involved in securing a data resource's long-term viability and subsequent use. They are not, however, essential to it. Provided that sufficient documentation exists, even data resources produced in idiosyncratic proprietary formats may, with sufficient investment, be emulated or rendered into platform independent (and thus migratable) formats
   2.3. The following categories of standards should be considered by data creators
      2.3.1. Those which ensure that data can be migrated with minimum content loss across platforms (standard file formats, and compression and encoding techniques)
      2.3.2. Those which ensure that data can be migrated meaningfully between individuals and organisations (documentation standards as devised by specialist communities and curatorial professions)
      2.3.3. Those which ensure that data resources are comparable with other like resources (data value standards as derived by specialist communities)
      2.3.4. Those which ensure that data resources suit the purposes for which they are created (as derived by specialist communities)

3. Data storage and data structures. These should be chosen to support intended use and/or preservation scenarios.
   3.1. In some instances, it may be appropriate to represent and store a data resource in different structures. Academic data archives, for example, may store data in the form in which they were deposited, in a form conducive to their long-term management, and in a number of forms as appropriate to their intended uses. Digitisers might similarly store data resources in differently structured versions, for example as high-quality master copies from which other lower-quality distribution copies may be generated as appropriate to different uses.
   3.2. Many stakeholders involved at some stages of a digital resource's life cycle may confront or need to implement remote data management strategies. These provide particular challenges particularly where data resources which are managed remotely need to be integrated from a users point of view with data and other information resources which are stored on site. Successful remote management typically entails agreement with third-party data managers about minimum level data selection, documentation, management, and delivery standards and practices.

4. Data documentation. Data documentation is essential to the exchange of data resources between platforms and individuals.
   4.1. At a minimum, data documentation should provide information about a resource's provenance, contents, structure, and about the terms and conditions attached to its subsequent management and use.
   4.2. At a minimum, documentation should be sufficiently detailed to support:
      4.2.1. Resource discovery (e.g. the location of a resource which is at least briefly described along with many other resources)
      4.2.2. Resource Evaluation (e.g. the process by which a user determines whether s/he requires access to that resource)
      4.2.3. Resource Ordering (e.g. that information which instructs a user about the terms and conditions attached to a resource and the processes or other means by which access to that resource may be acquired)
      4.2.4. Resource Use (e.g. that information which may be required by a user in order to access the resource's information content)
      4.2.5. Resource Management (e.g. administrative information essential to a resource's management as part of a broader collection and including information about location, version control, etc.)

   4.3. Where data resources are included or intended for inclusion within broader collections, minimum documentation should be supplied for all resources according to an appropriate standard or standards selected by the collection managers in light of their collection development aims, the kinds of data contained within their collection, and the information requirements of the collection's intended users. Collection managers should be advised of standard documentation practices.
   4.4. Those responsible for managing digital collections are typically best able to determine how data resources in that collection should be documented. Those responsible for creating resources within a collection are typically best able to supply information required for its adequate documentation. The documentation of data which is included in collections should result from a dialogue between data creators and data managers.

5. Preservation strategies
   5.1. Data migration is the preferred preservation strategy for data resources which are created with platform-independent data standards or which can be migrated into such data standards with minimal content loss. Such resources may be preserved by ensuring their readability on contemporary media and, where necessary, by reformatting them as required by ascendant standard regimes.
   5.2. Migration is an inappropriate strategy for the following kinds of resources:
      5.2.1. Those for which the hardware platform contributes makes an essential contribution to the resource's meaning and/or to the experience of its use (e.g. video games, game boys);

5.2.2. Those where data are stored in un-documented proprietary formats (proprietary information systems which store data in undocumented binary formats);
5.2.3. Those where data are stored in un-documented formats and bundled with access software which is also undocumented (commercial CD-ROM products)

For such resources, technical preservation (in the case of 1 and 3) or emulation (1-3) may be preferred.

The study uncovered little experience of either emulation or technical preservation. Such experience should be developed.

6. Preservation practices (with regard exclusively to data migration)
   6.1. Data preservation entails the creation and maintenance of archive copies of data files. Archive copies of a data resource are independent of any on-line representation and as such are distinguished from back-up copies of the same resource. Periodic and systematic back-up of on-line data resources is not itself a sufficient preservation strategy.
   6.2. Archive copies should be stored on industry standard digital tape or on other approved contemporary media as may arise.
   6.3. Archive copies should be available on- and off site. Off-site copies should be stored at a safe distance from on-site copies to ensure they are unaffected by any natural or man-made disaster affecting the on-site copies.
   6.4. Archive copies should be written with different software to protect data against corruption from malfunctioning or virus- or bug-ridden software
   6.5. Archive copies should be made to comparable magnetic media purchased from different suppliers to guard against faults introduced by the media's suppliers into their products or into batches of their products.
   6.6. Data files stored as archive copies should be migrated to new media. Migration should take place within the minimum time specified by the media's supplier's for the media's viability under prevailing climatic conditions. In addition, media should be checked periodically for their readability. Such checking may be conducted automatically by archive systems according to parameters set by system operators.
   6.7. The integrity of data files should be checked periodically using checksum and other like procedures. Such procedures may be implemented automatically by the archive system according to parameters set by system operators
   6.8. Proper preservation is expensive requiring substantial computing infrastructure and expertise not normally accessible to all those involved in the development and management of data collections, even as data archives. Those individuals and organisations which lack the appropriate facilities should be advised to conduct a cost benefit analysis and to determine whether the data preservation functions they require may be most cost effectively outsourced to a specialist computing service, data bank , or other organisation.

7. Data use
   7.1. Data creators' fear that their data may be put to unwarranted or inappropriate uses is a principal deterrent to the development of high-quality data collections. Robust and enforceable user agreements, combined with user registration, authentication, and other security measures will go someway toward alleviating this fear and enhancing collection development activities. Investigation into the development and widespread deployment of such mechanisms is seen as a priority consideration.
   7.2. Users and data developers alike show a growing preference for making data resources available over the Internet via World Wide Web browsers. Web delivery is an appropriate and cost effective means of delivering some resources. For others, it adds a significant development cost and may reduce a resource's functionality.

## 7. Bibliography, Resources and References

**Contents**

2. General resources and projects
3. Related resources
4. Bibligraphies and links
5. Studies and other related publications
6. References

### 7.1 Introduction

This section provides a guide to general WWW resources covering the creation and preservation of digital resources, and for selected sites providing information on individual digital preservation initiatives, issues, standards, technologies, bibliographies, and related topics. It is not intended as an exhaustive list so much as a high-level introduction to third-party directories and sites which between them will provide an exhaustive coverage of a particular topic or theme covered in the Study.

### 7.2 General Resources and Projects

Australian Archives: Managing Electronic Records
<http://www.aa.gov.au/AA_WWW/AA_Issues/ManagingER.html
Publication, incorporating guidelines and standards, documenting the Australian Commonwealth's efforts in preserving government and public documents through strategic management of electronic records; development and implementation of electronic record -keeping systems; and, migration of electronic records, with their content, structure and context intact, across changes in software and hardware platforms.

Commission on Preservation and Access (US) <http://www.clir.org/cpa/
The CPA aims "to foster, develop, and support collaboration among libraries and allied organisations to ensure the preservation of the published and documentary record in all formats, and to provide enduring access to scholarly information". Its web pages maintain a Newsletter, extensive reports on various aspects of preservation, pointers to technological initiatives of relevance to digital collections and digital libraries, and pointers to preservation science research.

Digital Library Federation (US) <http://www.lcweb.loc.gov/loc/ndlf/
A consortium of fifteen of the US's largest libraries and archives cooperating to ensure access to digitised materials. The site provides links to member organisations with a range of digital library access and preservation projects.

DLM-FORUM Electronic Records <http://www.echo.lu/dlm/en/home.html
The European Commission organises " a multidisciplinary Forum to be held in the framework of the Community on the problems of the management, storage, conservation and retrieval of machine-readable data, inviting public administrations and national archives services, as well as representatives of industry and of research, to take part in the Forum". Proceeding of the DLM-FORUM in December 1996 are available in hard-copy published format (European Commission 1997a) or can be downloaded as PDF files from <http://www.echo.lu/dlm/en/proc-index.html. An updated and enlarged edition of the Guidelines on best practices for using electronic information prepared for the conference is also now available as a hard-copy publication (European Commission 1997 b) or can be downloaded as PDF files from <http://www.echo.lu/dlm/en/gdlines.html.

European Preservation Information Centre (EPIC) <http://www.knaw.nl/ecpa
EPIC's web pages provide a "gateway for information on work directed at the preservation of the documentary heritage in Europe: first of all paper-based materials, but also sound, film, photographs, and digital archives".

International Council of Archives Committee on Electronic Records
<http://www.archives.ca/ica/english.html#top
The ICA is a non-governmental organisation founded in 1948 at a meeting convened by UNESCO. With some 1,350 members in over 150 countries, it acts internationally "to promote

the preservation, development and use of the world's archival heritage. The ICA's Committee on Electronic Records was established in 1993 "to undertake study and research, promote the exchange of experience and draft standards and directives concerning the creation and archival processing of electronic records". Currently, the Committee's web site hosts its Guide for Managing Electronic Records from an Archival Perspective, its Electronic Records Programs: Report on the 1994/95 Survey, and its Electronic Records: Literature Review.

Inter-university Consortium for Political and Social Research (ICPSR, US) <http://www.icpsr.umich.edu/ICPSR/
The Inter-university Consortium for Political and Social Research is an organisation of member institutions working together to promote and facilitate research and instruction in the social sciences and related areas by acquiring, developing, archiving , and disseminating data and documentation for instruction and research, and by conducting related instructional programmes. A strategic undertaking of the ICPSR is the acquisition and long-term preservation of social science data. There is a wide range o f informed and detailed information on the site on digital archiving of social science data built on many years experience in this field.

International Federation of Library Associations (IFLA) <http://www.nlc-bnc.ca/ifla/home.html
IFLA is "a worldwide, independent organisation created to provide librarians around the world with a forum for exchanging ideas, promoting international cooperation, research and development in all fields of library activity". IFLA's Core Programme for Preservation and Conservation focuses "efforts on issues of preservation and initiate(s) worldwide cooperation for the preservation of library materials". Its web pages provide extensive listings of Electronic Collections and Services and Digital Libraries: Resources and Projects.

International Standards Organisation (ISO), Open Archival Information (OAIS) Standard <http://bolero.gsfc.nasa.gov/nost/isoas/overview.html
ISO is working to develop standards for the long-term preservation of digital information obtained from observations of the terrestrial and space environments and which could also apply to other long-term digital archives. ISO aims to provide a framework and common terminology that may be used by Government and Commercial sectors in the request and provision of digital archive services. Working papers and drafts (White Books) of the Reference Model for an Open Archival Information System are available on the web and can be downloaded by ftp.

Internet Archive (US) <http://www.archive.org/
The Internet Archive is collecting and storing public materials from the Internet such as the World Wide Web, Netnews, and downloadable software which have been donated by Alexa Internet. The Archive will provide historians, researchers, scholars, and others access to this vast collection of data (reaching ten terabytes), and ensure the longevity of this information.

Library of Congress (US) <http://www.loc.gov/
"The Library of Congress is committed to assuring long-term, uninterrupted access to the intellectual content of the Library's collections, either in original or reformatted form. This mission is accomplished directly through the provision of conservation, binding and repair, reformatting, materials testing, and staff and user education; and indirectly through coordinating and overseeing all Library-wide activities relating to the preservation and physical protection of Library material". See Preservation (Library of Congress) and Library of Congress Preservation Directorate on this site for references to digital preservation.

National Archives and Records Administration, Center for Electronic Records (US) <http://www.nara.gov/nara/electronic/homensx.html
"Identifying permanently valuable records for retention in the National Archives involves co-operation between the National Archives and Records Administration (NARA) and the various agencies of the Federal Government. Through the process of appraisal, the Center for Electronic Records identifies and selects the electronic records it judges to have continuing value." The site provides extensive information on holdings and services, and information for record managers including "Managing Electronic Records - An Instructional Guide" <gopher://gopher.nara.gov/00/manager/federal/publicat/elcrecs

National Library of Australia <http://www.nla.gov.au/
Provides information on the National Libraries holdings and activities in electronic access and preservation. This includes guidelines and workplans for the Selection Committee on Online Australian Publications (SCOAP) and information on the PANDORA Project which is archiving a selection of Australian electronic journals.

National Library of Canada Electronic Publications Pilot Project <
http://collection.nlc-bnc.ca/e-coll-e/erereport.htm
From June 1994 to July 1995, the National Library of Canada (NLC) conducted a pilot project to acquire, catalogue, preserve and provide access to a small number of Canadian electronic journals and other representative publications available on the Internet. This project was called the Electronic Publications Pilot Project (EPPP). The Summary of the Final Report is available on-line. The Electronic Publications Pilot Project (EPPP) Final Report (June 1995) is available as a downloadable file.

Natural Environment Research Council <http://www.nerc.ac.uk/environmental-data/
The Natural Environment Research Council (NERC)is the lead body in the UK for funding research, survey, monitoring and training in the environmental sciences. NERC plays a key role in environmental data management. The site provides links to the Data Centres funded by NERC for long-term preservation of electronic data. The NERC developed an excellent Data Policy (NERC 1996) which has recently been updated (NERC 1998) and is also available from the NERC site.

Preserving Access to Digital Information (PADI, Australia)
<http://www.nla.gov.au/dnc/tf2001/padi/padi.html
Jointly sponsored by the Australian Archives, the Australian Council of Libraries and Information Services (ACLIS), the National Preservation Office and the National Film and Sound Archive PADI preserves "access to significant Australian information in digital form [by] providing mechanisms that will help to ensure that digital information is managed with appropriate consideration for preservation and future access." Particular objectives include the development of national strategies, a Web site for information and promotion purposes, a forum to exchanging views, preservation and access guidelines and strategies, and to impact on preservation aspects of national digital information initiatives, and actively identify and promote relevant activities. PADI's web site contains "selected bibliographies on preserving access to digital information. The bibliographies... are selected from the many available. They include material that is both on-line and in paper-based formats".

Public Record Office (PRO, UK)
<http://www.pro.gov.uk/http://www.open.gov.uk/pro/prohome.htm
The PRO has established two digital preservation programmes, EROS (Electronic Records from Office Systems) for the records created in the office networks being introduced by government bodies, and NDAD (UK National Digital Archive of Datasets at <http://ndad.ulcc.ac.uk) for structured datasets, such as survey files and databases, in government departments.
Final Guidelines on the Management of Electronic Records from Office Systems have been published <http://www.pro.gov.uk/eros/guidelinesfinal.pdf.

Research Libraries Group (RLG, US) <http://www.rlg.org/
"A not-for-profit US corporation of institutions devoted to improving access to information that supports research and learning". Its Preservation Service - PRESERV program "is a series of projects and activities that support local institutions in their efforts to preserve and thereby improve access to endangered research materials. Current projects include digital archiving, preservation and reformatting, digital image capture, preservation issues of metadata and preserving magnetic media. PRESERV focuses on achieving collaborative and innovative solutions to common preservation problems by inventing and testing new models for cooperation.... Working Groups established to achieve PRESERV's Strategic Plan include those on Digital Archiving, Preservation and Reformatting Information, Digital Image Capture, Preservation Issues of Metadata, and Preserving Magnetic Media".

Time and Bits <http://www.gii.getty.edu/timeandbits/index.html

An integrated technical and philosophical discussion of digital archives and their future that includes the socio-cultural and economic implications of both the problems and the solutions that could provide a framework for long-term digital cultural preservation sponsored by The Getty Conservation Institute, the Getty Information Institute, and the Long Now Foundation [San Francisco]. In February 1998, a small group was convened at The Getty Center to share concerns and expertise in technology, culture , and time. This Web site provides the background papers for that discussion, proceedings, press coverage, links to other resources and projects, and ongoing moderated discussion.

Universal Preservation Format (UPF, US) <http://info.wgbh.org/upf/
Sponsored by the WGBH Educational Foundation and funded in part by a grant from the US National Historical Publications and Records Commission of the National Archives, the Universal Preservation Format Initiative advocates a format for the long-term storage of electronically generated audio-visual and other media. This site disseminates information about the proposed universal preservation format and related initiatives. One of the links on the site is the UPF Needs Assessment Survey, devised to assist in incorporating the needs of potential users into the platform-independent UPF.

**7.3 Related Resources**

British Standards Institute <http://www.bsi.org.uk/
Standards for buildings, computer hardware and software, data storage, and related matters. Also BSI PD00008 covering legal admissibility of electronic information (references BSI 1996).

CD Information Center (US) <http://www.cd-info.com/
A central reference point to learn the basics about Compact Disc development and production technology and the industry.

Central Computing and Telecommunications Agency (CCTA, UK) <http://www.ccta.gov.uk/
CCTA is the Central Computer and Telecommunications Agency, an agency of the UK government's Office of Public Service. To help Business Managers CCTA has written many publications on IT management and offers consultancy to support its published guidance.

Computer Conservation Society (CCS, UK) <http://www.cs.man.ac.uk/CCS/
The Computer Conservation Society was formed in 1989 as an initiative between the British Computer Society and the Science Museum of London. The aims of the CCS are to: promote the conservation of historic computers; develop awareness of the importance of historic computers; and encourage research on historic computers. The site provides links to related sites and details of the Society's working parties, lecture programme, and its bulletin "Resurrection".

Information Infrastructure Standards and Specifications
<http://nii.nist.gov/ext_links/standards.html
Resources compiled as part of the US National Information Infrastructure Initiative, providing access to information on a wide range of standards.

Information 2000- OII Service <http://www.echo.lu/oii/en/oii-info.html
The Open Information Interchange (OII) Service started as part of the European Commission DG XIII's IMPACT 2 programme and is continuing under INFO 2000, the Commission's multi-annual programme for stimulating the development of the European multimedia content industry and encouraging the use of multimedia content in the emerging information society. The focus point of the OII initiative is the comprehensive listing of published standards provided in the OII Standards and Specifications List <http://www2.echo.lu/oii/en/oiistand.html#oiistand/. This list is wide-ranging and a valuable pointer to: sectoral standards for museum information, library information, archiving, etc.; electronic commerce; data coding; communications; and related areas. The list is complemented by a series of OII Guides, the OII Fora List, reports and a comprehensive index.

National Institute of Standards and Technology- Threat Assessment
<http://csrc.ncsl.nist.gov/nistir/threats/

"Threat assessment of Malicious Code and Human Threats" - The NIST assessment of the history of 'malicious code' (viruses and worms) includes a discussion of 'malicious code', summarises protection methods, and projects future threats from both malicious code and human attack.

National Media Laboratory (US) <http://www.nta.org/AboutNTA/AboutNML/
The National Media Laboratory (NML) is an industry resource supporting the U.S. Government in the evaluation, development, and deployment of advanced storage media and systems. NML endeavours to provide a broad perspective of current progress in information technology issues, both from a commercial and a government perspective. The Web site is a valuable guide to current digital storage media and systems and research on their longevity as archival media.

Stanford University Libraries' Conservation OnLine <http://www.palimpsest.stanford.edu/
"A project of the Preservation Department of Stanford University Libraries, a full text library of conservation information, covering a wide spectrum of topics of interest to those involved with the conservation of library, archives and museum material s". Its web pages provide a bibliography of resources and projects under several headings relevant to digital media including bibliographies and resources guides, digital images, documentation, electronic media, electronic resources, etc.

Webreference: watermarks <http://www.webreference.com/multimedia/watermarks.html
Links to commercial companies and articles on watermarking of digital images.

### 7.4 Bibliographies and Links

Bibliography of Materials Relating to the Preservation of New Technology and Preservation Using New Technology (AU) <http://www.nla.gov.au/1/pres/pubs/bibmj.html
National Library of Australia's resources list on electronic preservation and technology issues.

Bilderbank (CH) <http://foto.chemie.unibas.ch/bilderbank/links/preserve.html
A Swiss photography and imaging research consortium of academic and commercial organisations, which maintains link pages for Preserving Digital Information, and Digital Watermarking.

Image and Multimedia Database Resources <http://sunsite.berkeley.edu/Imaging/Databases/
Howard Besser's and Rachel Onuf's comprehensive resource guide with sections on issues of relevance to digital preservation.

Introduction to Bibliography of Standards and Selected References
<<http://www.nlc-bnc.ca/resource/presv/einto.htm
Resource list relating to preservation in libraries compiled by Suzanne Dodson and Johanna Wellheiser of the National Library of Canada.

Partnership Potential: The Archives and the Data Archives. A Bibliography of Selected Works
<http://datalib.library.ualberta.ca/iassist/bib.html
An excellent "introduction to literature of the past three decades that represents the work of the international social science data community.

Preservation of electronic information: a bibliography
<http://www.ukoln.ac.uk/~lismd/preservation.html
Information resources page maintained by Michael Day, UK Office for Library and Information Networking (UKOLN).

Readings on Digital Libraries <http://www.sis.pii.edu/~lis2002/diglib-readings/
A bibliography of materials relating to digital information, libraries, archives, systems and related areas.

## 7.5 Studies and other related Publications

Berkeley Digital Library's Collection Policy <http://sunsite.berkeley.edu/Admin/collection.html

A succinct example of a collections policy developed for a digital library with a defined hierarchy of collection levels for digital library materials.

"Digital Preservation: A Time Bomb for Digital Libraries"
<<http://www.uky.edu/~kiernan/DL/hedstrom.html
Article by Margaret Hedstrom discussing digital preservation requirements, limitations of current preservation strategies, storage media, migration, conversion, management tools and related issues.

Joint Study on Digital Preservation <http://www.clir.org/cpa/reports/joint/
A collaborative study by Cornell University, Xerox and the Commission on Preservation and Access investigating the digital preservation of library materials.

"Long Term Intellectual Preservation" <http://aultnis.rutgers.edu/texts/dps.html
Discusses the impact and relationship of electronic information and its preservation on "intellectual preservation" or authenticity.

"Intellectual Preservation: Electronic Preservation of the Third Kind"
<http://aultnis.rutgers.edu/texts/cpaintpres.html
Discusses what the author considers the third dimension in the long-term digital preservation problem: intellectual preservation, or, authenticity and integrity.

Oxford University Policy on Computer Archiving Services
<http://info.ox.ac.uk/oucs/services/archiving/archive-policy.html
A useful overview of the introduction of the Hierarchical File Server at Oxford and the digital archiving policy established for the University.

PANDORA -Boxing for survival: Archiving, preservation and access issues related to Australian Internet based publications <http://www.nla.gov.au/nla/staffpaper/wsmith3.html
An excellent overview of the preservation and access issues relating to Australian Internet-based publications.

Preserving Digital Information: Final Report and Recommendations
<http://www..rlg.org/ArchTF/
Seminal report from the Task Force on Digital Archiving established by the Commission on Preservation and Access and the Research Libraries Group and published in 1996 (Garrett and Waters 1996)

Preserving the Electronic Assets of a University
<http://users.ox.ac.uk/~alex/hfs-AXIS-paper.html
Introduction to the IT Strategy of Oxford University and the issues and approach developed to preserving a growing range of electronic information generated by the University.

Reference Model for an Open Archival Information System
<http://bolero.gsfc.nasa.gov/nost/isoas/ref_model.html
Working drafts (White Books) of the Reference Model being developed by CCSD Panel2 of ISO to provide a framework and common terminology for specifying digital archive services.

Statement on the Preservation of Digitized Reproductions
<http://www.archivists.org/governance/resolutions/digitize.html
A statement of approved principles by the Society of American Archivists.

**7.6 References**

Beagrie, N, forthcoming, Developing a Policy Framework for Digital Preservation, (European Research Consortium for Informatics and Mathematics) Workshop Proceedings, Sixth Delos Workshop on Preservation of Digital Information, Tomar, Portugal, 17-19 June 1998.

BSI 1996, Code of Practice for Legal Admissibility of Information stored on Electronic

Document Management Systems, DISC PD 0008, British Standards Institute London.

European Commission 1997a, Proceedings of the DLM-Forum on electronic records, Brussels 18 to 20 December 1996, INSAR - European Archives News, Supplement II, 1997, Office for Official Publications of the European Communities Luxembourg.

European Commission 1997b, Guidelines on best practices for using electronic information, INSAR - European Archives News, Supplement III,1997, Office for Official Publications of the European Communities Luxembourg.

Garrett, J and Waters, D 1996, Preserving Digital Information. Report of the Task Force on Archiving of Digital Information commissioned by the Commission on Preservation and Access and the Research Libraries Group Inc.

Greenstein, D, 1997, Managing Digital Collections. Part I. Towards a Strategic Framework for the Development of Appropriate and Effective Organisational Data Policies, *The New Review of Information Networking*, 2(1997), 23-42.

Greenstein, D, 1998, Managing Digital Collections. Part II. In Search of Guidance, *International Journal of Electronic Library Research*, 1:4(1997), 433-54

Fresko, M 1996, Long Term Preservation of Electronic Materials. A JISC/British Library Workshop as part of the Electronic Libraries Programme (eLib). Organised by UKOLN 27th and 28th November 1995 at the University of Warwick, BL R&D Report 6328, The British Library.

Mackenzie Owen, J 1996, Preservation of Digital Materials for Libraries, Liber Quarterly, 6:4, 433-451

Mandel, C 1996, Enduring Access to Digital Information: Understanding the Challenge, Liber Quarterly, 6:4, 453-464

NERC 1996, Formal NERC Data Policy Statement, version 1.0 January 1996, Natural Environment Research Council Swindon.

NERC 1998, Formal NERC Data Policy Statement, version 2.0 February 1998, Natural Environment Research Council Swindon.

## Appendix 1. Guidelines for Digital Preservation: Draft Interview Questionnaire and Policy Framework

Version 2. 15/1/98
Neil Beagrie and Daniel Greenstein, Arts and Humanities Data Service Executive

**Contents**

1. Introduction
   1.1. Please give your name, your job title, and a brief explanation of your current role framing or implementing your organisation's data policy. Could you also briefly describe the nature of your collection and the reason it is being assembled ?

2. Validate the framework. The framework outlines the key strategic issues which need to be addressed in any data policy concerned with preservation and demonstrates the inter-relatedness between them.

   2.1. Does the framework cover the major strategic issues which must be addressed in any data policy concerned with preservation?

   2.2. Does the framework fully and correctly identify the inter-relatedness between these issues?

   2.3. How closely do the data flow and archive administration diagrams in the draft ISO Open Archival Information Systems Reference Model match your experience or expectations in digital archiving ?

3. Describe your organisation's data policy

   3.1. Data acquisition
      3.1.1. How do you decide what data resources to consider for inclusion in your collection?
      3.1.2. What criteria and procedures do you use to assess how, whetner and at what cost such resources can be included in your collection.
      3.1.3. What are your criteria and procedures for de-selecting resources from your collection?

   3.2. Data storage
      3.2.1. How are data in your collections structured (where structure refers to file formats, compression techniques, and encoding methods)?
      3.2.2. Why have you chosen these data structures?
      3.2.3. Under what circumstances would you re-structure your data?
      3.2.4. What data structure(s) do you request/require from depositors?

   3.3. Data documentation
      3.3.1. What documentation do you supply with any data resource and why?
      3.3.2. Can you supply examples of the data documentation that is supplied for data resources?
      3.3.3. Who supplies this documentation (e.g. depositors, in-house cataloguers)?
      3.3.4. What documentation do you request or require from data depositors?

   3.4. Data preservation
      3.4.1. Which of the following long-term preservation strategy or combination of strategies does your organisation pursue and why has it chosen to pursue those strategies?

         - Migration (data are stored in software-independent format and migrated through changing technical regimes)
         - Technology preservation (data are preserved along with the hardware and/or software on which they depend)
         - Emulation (the look, feel, and behaviour of a data resource is emulated on successive hardware/software generations
         - Other (please elaborate)

      3.4.2. How are preservation strategies implemented and what standards of accepted industry practices (if any) apply to them? We are particularly interested in your comments on:

3.4.3. What platforms and media do you use for storing data?
3.4.4. How many copies do you make of any data resource? In what data structures are those copies stored? When are copies made?
3.4.5. How, and how frequently do you refresh the media on which data resources or stored?
3.4.6. How and how frequently do you check the integrity (vaildate) of a data resource or the media on which it is stored?
3.4.7. What disaster recovery contingencies do you have in place?

3.5. Data use
3.5.1. How are data disclosed to end users (what catalogues exist; how and to whom are they made available)?
3.5.2. How are data supplied to end-users (what delivery media, data structures, delivery mechanisms do you supply)?

4. Rights management and legal issues
4.1. What is your organisation doing to protect the rights that are invested in the data deposited with it or to ensure its compliance with any legal obligations such as data protection ?
4.2. Would you supply copies of model licences etc.?

5. User support
5.1. Do you supply documentation/guidance for depositors (e.g. procedural guidelines for documenting, delivering, and licensing a resource) and could you supply a copy?
5.2. Do you supply documentation /guidance for users and could you supply a copy?
5.3. If your organisation could exercise some influence over the behaviour of data creators or data users, what would that influence be and why would you try to exert it? Does your organisation attempt to exercise some influence over data creators users?

6. Other aspects and issues
6.1. Are there other aspects of your archiving practice we should be aware of? If so what are they?
6.2. Are there any other organisations or individuals that can throw light on these issues?
6.3. Are there any publications or other documents that may shed some light on these issues?

# Guidelines for Digital Preservation: Draft Data Policy Framework Document

Version 2, 8/12/97
Neil Beagrie and Daniel Greenstein, Arts and Humanities Data Service Executive

**Contents**


## Introduction

Computerisation is changing forever the way information is being created, managed and accessed and is revolutionising our ability to communicate, analyse and re-use that information. At the same time electronic data needs active management from its creation if it is to survive and be kept accessible in a technological environment where there is rapid change and evolution of hardware and software. These factors have led to increasing concern about and research into the long-term preservation of digital information.

Substantial digital preservation initiatives are currently underway in Britain, for example at the British Library, the Public Record Office, the Data Archive, the National Environmental

Research Council, and the Arts and Humanities Data Service. Further initiatives are contemplated by the Joint Information Systems Committee, by the British Library, and by individual heritage and educational agencies which find themselves increasingly concerned with long-term preservation of the digital information resources which they are helping to create or archive. Growing British interest in digital preservation is complemented and shared internationally for example by the work of the Commission on Preservation and Access and the Research Libraries Group in the US; by the National Library of Australia; and by various initiatives in Europe and elsewhere.

To manage and preserve the digital components of our intellectual and cultural heritage, some high-level agreement is required about the application of common standards and best practice. At the same time, data archivists and scholarly and cultural heritage organisations which are developing digital collections, can benefit from sharing experience and guidelines for best practice in their preservation.

A key part of this shared experience and guidance is recognition of the importance of developing data and collection policies which reflect the whole life cycle of digital resources and the complex inter-relationships between different practices which may be adopted to create, use or preserve them. Digital preservation is crucial as part of a series of other issues which effect the creation, storage and use of a resource. These issues are all interdependent and suggest the need for an integrated data policy when approaching resource creation and management.

## Aims

This study aims to identify current practice, strategies and literature relating to the preservation of digital collections and to prepare and validate guidance for those wishing to develop collection or data policies for digital materials.

To scope and validate guidance on collection or data policies we are preparing an outline framework into which the results of our study will be integrated. Data preservation options constrain and are highly constrained by how data is created, what data is acquired, how they are documented and stored, and how the are intended to be used. Accordingly, where long-term preservation is a priority for collection managers, they will normally begin by defining their preferred preservation strategy and let this constrain other decisions which need to be taken with regard to the development, management and use of a particular collection.

The framework will be an outline data policy in which we hope to summarise best practice as it exists amongst data archivists and data managers who have experience of preserving digital information resources. It is based upon the life cycle of a digital resource and upon our understanding of how practices adopted at any stage ramify forward and back for others. A brief review of the stages and their intrinsic inter-relatedness is provided below.

## Contents

### 1. Data creation

Data creation entails a range of decisions which determine a data resource's intellectual content, structure, format, compression, encoding, and the nature and level of descriptive information. Accordingly how data is created will impinge directly upon how it can be managed, used and preserved at any future date. Decisions on data creation may be taken for a number of purposes and may often be beyond a curator/archivists control. Curators interested in the long-term preservation of the data must therefore often develop a dialogue with data creators and consider the implications of how a resource has been created and documented for its preservation.

### 2. Data Acquisition

Acquisition involves decisions about collection policy and which data resources should be selected for long-term preservation and included in (or excluded from) a collection. It will also involve data evaluation - a nuts and bolts assessment of those data resources which are potential

acquisitions and will determine how (even whether), and at what cost a data resource may be included in a collection. Decisions taken here will shape the collection and impinge directly upon how it is catalogued and documented, managed, made accessible to end users, and preserved.

**3. Data management**

A suite of related decisions about how data resources are handled and described once they are included in a collection. How data is managed will depend upon how it has been created or supplied (e.g. in what format, with what documentation, and under what terms and conditions). Data management options will accordingly be constrained by decisions taken when data is created or selected for inclusion in a collection. They will also constrain data use and preservation options. The suite of decisions are outlined below in greater detail:

1. Data structure, format, compression, and encoding. How data is formatted (written to magnetic media), compressed, and encoded (i.e. how internal semantic or syntactic features are represented) will determine its portability across hardware and software platforms and how it may be stored, manipulated, and subsequently enriched.
2. Data documentation. The information supplied about a data resource's structure, contents, provenance, and history. It influences how a resource is located, managed, and used, and frequently reflects data acquisition decisions (notably as they reflect what documentation is supplied for a resource, how it is supplied, and who supplies it).
3. Data storage. Involves organisational decisions about whether collections or parts of collections are stored centrally or distributed across several sites, and technical decisions about what magnetic media and hardware platforms are used. Options are constrained by resources' structure, but also upon the relative emphasis given by a collection to their use and/or preservation. Accordingly data storage decisions will constrain data creation or accessioning ones and help to determine how (even whether) and to what extent a data resource once included in a collection can be preserved and/or used.
4. Data preservation. A suite of strategic and procedural decisions which together help to ensure that a data resource survives through time and changing technologies with minimal loss in its information content, functionality, and accessibility. Decisions involve the adoption of a preservation strategy or combination of strategies normally taken from the following list:
    - Migration (data is stored in software-independent format and migrated through changing technical regimes)
    - Technology preservation (data is preserved along with the hardware and/or software on which it depends)
    - Emulation (the look, feel, and behaviour of a data resource is emulated on successive hardware/software generations)
5. Decisions are also required about how to protect data integrity in the short as well as in the longer term. Consideration will normally given to:
    - periodic assessment (validation) of a resource's its completeness, function, and consistency;
    - copying a resource (possibly onto alternative media and into alternative formats);
    - periodically moving a resource from older to newer or "fresher" storage media.

**3.5 Data use**

Since data policies which facilitate access are not always those which ensure cost effective long-term preservation, the tension between access and preservation requirements are most acutely felt at this stage. How data is delivered to and used by end users will be contingent upon how and why it was created or acquired, how/where it is stored , and upon what software and hardware is needed to access it. Because use is so highly contingent, many collections managers construct data policies by identifying preferred user scenarios and allowing these to constrain decisions about collection development, management, and data preservation.

**3.6 Rights management and legal issues**

Not a stage in the life cycle of a digital resource so much as a consideration of intellectual and property rights, and of related legal issues including data protection and confidentiality, or the legal obligation to select and preserve categories of records, which need to be made at every

stage. What rights are vested in a resource will impinge on how and whether it may be represented in machine-readable form; how, by whom, and under what conditions it may be used; how it can and should be documented and even stored (e.g. where 'sensitive' information requires encryption); and how, whether, and by whom it can legally be preserved.

## Draft ISO Open Archival Information System (OAIS) Standard

ISO is working to develop OAIS standards for the long term preservation of digital information obtained from observations of the terrestrial and space environments. ISO aims to provide a framework and common terminology that may be used by Government and Commercial sectors in the request and provision of digital archive services.

We intend to explore the compatibility and relevance of the work undertaken by the ISO OAIS working party on behalf of the Space and Earth Observation communities to our own study and other digital archives. Draft data flow and archive administration diagrams from the ISO OAIS Reference Model (White Book version 2.0) are given below. Further information and definition of the terms are available from the OAIS Reference Model web page.