Developing a Policy Framework for Digital Preservation

Neil Beagrie
Arts and Humanities Data Service Executive
King's College London
Strand
London WC2R 2LS

Abstract

The Arts and Humanities Data Service (AHDS) has been established by the Joint Information Systems Committee of the UK's Higher Education Funding Councils to collect, preserve and promote re-use of digital resources which result from or support research and teaching in the arts and humanities.

Within the UK, the Digital Archiving Working Group (DAWG) has been formed to co-ordinate research into digital archiving between its members: the British Library, the National Preservation Office, the Higher Education sector, the Public Record Office, the Research Libraries Group, and the Publishers Association. DAWG has commissioned a series of studies to examine issues raised in the CPA/RLG Task Force report on Digital Preservation.

As part of this research programme the Executive of the AHDS has recently completed a study into developing a strategic policy framework and implementation guidance for the creation and preservation of digital resources. The framework is based on the life-cycle of a resource from its creation, management and preservation through to use, and examines the dependencies and issues at each stage, how these are interlinked, and the influence of the legal and business environment. The framework and the implementation guidance have been based on interviews with representatives of those organisations which have a major stake in the creation or long-term maintenance of digital research and on an extensive literature survey. Institutions interviewed included a selection of international libraries, museums, archives, scientific and other academic data archives. Together they represent a synthesis of work to date on digital preservation and point directions toward the development of robust preservation strategies.

This paper presents a summary of the major findings and the policy framework from the study report.

1. Introduction

1.1 The Importance of Preservation and Access to 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.

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. 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.

1.2 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.

1.3 The DAWG 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. The Joint Information Systems Committee subsequently agreed to fund a research programme implementing the recommendations, to be guided by the Digital Archiving Working Group (DAWG) in the UK and administered by the British Library Research and Innovation Centre.

1.4 Aims of the AHDS Study

The study undertaken by the AHDS Executive (Beagrie and Greenstein forthcoming) as part of the DAWG Programme of Preservation Studies, 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.

The 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.

1.5 Methodology

The study was carried out by the author 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 took account of the policies and programmes which large-scale digital preservation and digital collection development initiatives are beginning to provide in some "published" format.

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.

2. The Policy Framework

2.1 The Development of the Policy Framework

The starting point for the study was a 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 Higher Education Funding Councils in the UK.

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. Further information on the AHDS and the AHDS collections policy is available from the AHDS website at <http://ahds.ac.uk/>.

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 (eg 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.

2.2 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 included in the study report 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 in the report.

2.3 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 focussed 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.

2.4 Legal and Economic Environment

This is 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.

2.5 The Life-Cycle of the Resource

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.

Data and Collection Management and Preservation

Data and collection management and preservation may involve a number of stakeholders who can fulfil 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 m ay 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.

Data management and preservation 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 to 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.

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.

Data Use

Data use can occur immediately after its creation and for an indefinite period thereafter. Its use can be to fulfil its primary purpose when created, involve subsequent secondary analysis, or inclusion in a collection developed to fulfil 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.

3. Conclusions

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 digitaldata 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 the AHDS 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 and 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 digitalresources 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 in the study. 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. Organisations which retain digital information to document their activities and for other purposes, may have the same perspective and the same degree of control.

Acknowledgements

I am extremely gratefully to Daniel Greenstein for his input and comments on an earlier draft and to the individuals who were interviewed and contributed substantially to the study report.

References

Beagrie, N and Greenstein, D forthcoming, Digital Collections: a strategic policy framework for creating and preserving digital resources.
<URL for consultation draft: http://ahds.ac.uk/manage/framework.htm>

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.<URL:http://www.echo.lu/dlm/en/proc-index.html>

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.
<URL: http://www.echo.lu/dlm/en/gdlines.html>

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. Commission on Preservation and Access, Washington DC.
<URL:http:www.rlg.org/ArchTF/>

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 Research and Innovation Centre, London.
<URL:http://www.ukoln.ac.uk/services/papers/bl/rdr6238/>