# 3. ENCOURAGING AND LEARNING FROM BEST PRACTICE

## 3.1 Different stakeholders, different interests

While the previous chapter examined the different views of the various stakeholders involved in digital information, this chapter takes another perspective and looks at the subject from the point of view of the life-cycle of the digital object itself. Different stakeholders become involved with data resources at different stages. Indeed, few organisations or individuals that contribute to the development and 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, but 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.

The result is that digital preservation is essentially a distributed process, which engages a range of different (and often differently interested) stakeholders who become involved with digital resources at particular phases of their life-cycle. Although stakeholders have a clear understanding of their own involvement in digital resources, they have less understanding of the interests of others. Further, they may have little or no understanding of how their own involvement influences (or is influenced by) others, nor awareness of the current challenges in ensuring the long-term preservation of the cultural and intellectual heritage in digital form. To increase the prospects for digital preservation—and reduce its cost—different groups of stakeholders need to become more aware of how their particular involvement with a digital resource ramifies across its life-cycle.

In other words, we need to look at ‘the big picture’.

## 3.2 Experience, cross-fertilisation and information-sharing

Some groups of organisations, including data banks, institutional archives and academic data archives, have long experience of managing data over the longer term. The library and cultural heritage sectors have initiated further research and development in this area. Despite their different aims, and the different business, funding and legal environments in which they work, these stakeholders have a great deal in common; some have 30 years and more of highly relevant data management experience. Nonetheless, there are few channels to facilitate their inter-communication. Cross-fertilisation and information sharing are crucial to the success of long-term preservation. 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.

The challenges posed by digital information are leading to a 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. As the digital store grows and diversifies we become more aware of the need for selection, standards and cooperation between different organisations.

All stakeholders acknowledge as essential the use of standards throughout the life-cycle of a digital resource. Their application ensures that data resources fulfil at minimum cost the objectives for which they were made. They also facilitate and reduce the cost of interchange across platforms and between individuals and institutions. The selection and use of standards, however, is highly contingent upon where in its life course any individual or organisation encounters a digital resource, and on the role which that individual or organisation plays in the creation, management, or distribution and use of the resource.

Little 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. We need to identify and document such ‘best practices’ in order to provide integrated access to them in a meaningful way.

The complex inter-relationships between different practices involved in the life-cycle of digital resources have suggested the need for an integrated policy framework to develop a cost-effective approach to resource creation, preservation and use.

### 3.3 An integrated policy framework

The Arts and Humanities Data Service (AHDS) has developed a policy framework based on the concept of the life-cycle of a digital resource, with the aim of assisting those involved in the creation of and access to digital resources in identifying and addressing key issues and in developing their own data policies (Beagrie and Greenstein, 1998).

The framework outlines the main stages 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 against the implications for preservation of those resources designated of long-term cultural and intellectual value. It provides a high-level checklist which individuals and institutions can use to develop policies and guidance which they will tailor to their specific needs. 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.

### 3.4 The life-cycle of a digital resource

Table 3 summarises the main stages in the life-cycle of a digital resource.

*Table 3. Life-cycle of a digital resource*

| | | |
|---|---|---|
| **1** | **Data creation** | Decisions made when the digital resource was created—often outside the control of the collection manager, but having a major impact on the options subsequently available. |
| **2** | **Data management and preservation** | |
| | Acquisition, retention or disposal | Decisions based on the digital resource's content, usability and relevance to users; the ease with which it can be managed, catalogued, made accessible and preserved. |
| | Data structure | How a digital resource is formatted, compressed and encoded. |
| | Data description and documentation | The extent to which the digital resource's structure, content, provenance and history have been documented. |
| | Data storage | The computer hardware and media used to store the digital resource. |
| | Data preservation | Safeguarding the information content of any digital resource from the ravages of time, technological change and decaying magnetic media. Different strategies are appropriate for different data types and structures. Preservation requirements will impinge on how digital resources are structured, documented, stored and validated and possibly even on the conditions and methods by which digital resources can be accessed by end users. |
| 3 | **Data use** | Decisions on how digital resources are to be delivered and used; will be influenced by how they were created and will influence how they are managed. |
| 4 | **Rights management** | Intellectual property rights, data protection and confidentiality issues; need to develop both acquisition licences and distribution licences and implementation procedures. |

All the key issues and all the elements of the framework are closely interrelated. Decisions about whether to create or include a digital resource in a collection—and about its content and format—will impinge on how it can be managed and stored, on how or even whether it can be preserved, and on how copies can be delivered to end users. Equally, the uses intended for a particular resource, or the method chosen to preserve it over time, should influence decisions taken when creating or including a digital resource in a collection.

The legal and economic environment surrounding the resource, interlinked with the organisational mission of its stakeholders, will also impact on the application of the framework. For example, the legal or contractual rights 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; and how, whether, and by whom it can legally be preserved.

Similarly, resources created in a commercial environment may have commercial constraints which can impinge on data management, preservation, and use, while the priorities and objectives of funding, and the funding agencies, for the resource throughout its life-cycle may also impact in a number of different ways.

Some organisations have begun to implement proactive strategies to influence and manage the life-cycle of digital resources. ‘Remote management’—initiatives taken to manage ‘active’ or ‘dynamic’ resources or contract for specialist skills and facilities—appears to be a widespread response to a distributed process and best practice in its use should be developed and encouraged.

### 3.5 The framework in practice

The AHDS framework is supported by a number of case studies which provide a synthesis of existing practice, policies and implementation strategies. They introduce a range of stakeholders and organisational roles in the creation, management and preservation of digital resources.

**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’ in 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 the future. The data bank ensures that deposited data are available on contemporary storage 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. Examples of data banks include the Oxford University Computing Service (OUCS), which provides an archive for the electronic assets of the University of Oxford, and the University of London Computing Centre (ULCC), which acts as a data bank for a variety of depositors and offers a data bank facility for the UK Public Record Office’s Computer Readable Data Archive.

**‘Digitisers’** create data resources, or build collections of resources which are either created or acquired from third parties, for a variety of different but always very specific purposes. Space missions which install satellites for the purposes of transmitting digital images of space, archaeologists who build a simulated town plan of Pompeii, art curators who hang a virtual exhibition, and librarians who digitise images of printed books are all ‘digitisers’. 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. The digitisers may be grouped in three broad categories which reflect their roles and their intentions 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.
- Library, archive, and cultural heritage organisations 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 on five 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 fragile objects within the collection). It is likely that the organisational missions of this group will develop over time as the balance of collections moves 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 that 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. They are increasingly interested in exploiting the value of their back files.

Examples of digitisers include: among the research organisations, the Space Data Centre (SDC) at the Rutherford Appleton Laboratory; and, among the cultural heritage organisations and libraries, the British Film Institute (BFI), the National Museum of Science and Industry, and the Victoria and Albert Museum (V&A).

**Funding and other agencies** invest in the creation of digital information resources and sometimes 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-cycle. Examples include the Natural Environment Research Council (NERC) and the Scottish Cultural Resources Access Network (SCRAN).

**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. Examples include 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).

**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. Examples include the Data Archive, at the University of Essex, and the Arts and Humanities Data Service (AHDS).

**Legal deposit libraries** have an 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. Legal deposit 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, and their orientation toward long-term preservation and secondary use. Examples include most national libraries.

## 3.6 Two case studies

Two organisations with interests in digital preservation are described here as examples of how key issues have been approached in practice and how different organisational missions shape approaches to the creation and preservation of digital resources.

The University of London Computing Centre (ULCC) fulfils the core functions of a **data bank**. It also 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. As a data bank, 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 file transfer protocol (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.

*Table 4: A data bank: University of London Computing Centre (ULCC)*

| STAGE IN LIFE-CYCLE | ULCC |
|---|---|
| **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 | This unique perspective is apparent in the core services offered at ULCC which accessions and stores data created in a variety of different standard and non-standard formats. Where ULCC's work with the PRO is concerned, PRO guidelines pertaining to the management of computer-readable datasets mitigate to a |

| | |
|---|---|
| in how, why, or for whom deposited data are created. | 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 and so its role in data selection is limited. | ULCC departs some way from this norm in its work for the PRO. Although 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 are responsible for identifying and preparing records for long-term retention. |
| **Data structure and storage** | |
| 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. They are therefore largely unconstrained in the data structures they can accommodate 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. | ULCC will undertake these additional functions when 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. Government departments take account of data resources' physical and technical characteristics when selecting data for deposit. ULCC will also restructure data deposited by the PRO since it is engaged to migrate them through changing technical regimes and make them accessible to users. |
| **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, but content migration (ensuring that data can be meaningfully represented by and accessed from contemporary platforms) is the responsibility of the depositor. The data bank 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 online representation they may have. | *A preservation scenario*:<br>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 similar 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 re-use. 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. | 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 services are also complicated by the data bank's involvement in 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. | Depositors take full responsibility for data they deposit in the archive. |

As a **funding agency**, the Natural Environment Research Council (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 value for scholarly re-use. 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 and act as repositories (academic data archives) for data created with NERC funding.

*Table 5. A funding agency: the Natural Environment Research Council (NERC)*

| STAGE IN LIFE-CYCLE | NERC |
|---|---|
| **Data creation** | |
| 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. They are 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- | NERC has adopted data policies which determine the life-course of grant-funded data from their inception, through to their creation, management, and subsequent use. NERC expects of its grant applicants the rigorous evaluation of both content and technical criteria that is conducted by other digitisers when planning a data creation initiative. 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 |

| | |
|---|---|
| usable; and (d) manageable over the longer term.<br>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.<br>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. | centre over the longer term.<br>In terms of content and technical criteria, NERC cannot be too prescriptive since it is funding research in a wide range of scientific disciplines.<br>Because it funds the development of a wide range of data resources which are created for very different purposes, NERC is not prescriptive, relying upon specialists involved in the application review process to advise about appropriate use of standards and best practice. |
| **Data management and preservation** | |
| Funding agencies take a serious interest in how data are managed and preserved because they recognise the long-term scholarly value of the research resources created by their grantholders. | 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 NERC may require from its grant holders are designed and implemented by the data centres with secondary users' 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, 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. |
| **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 advantages which may accrue from the collection, professional management, and re-use of such resources. | For NERC, both roles are undertaken through the data centres, though in conformance with 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, reflecting their diverse holdings and the different specialist user communities that each centre serves. In these respects, the data centres resemble academic data archives. 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. |

| **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. | The intellectual property vested in any data resources that are created by NERC employees are owned by 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 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 licence to distribute them for educational use. NERC also takes pains through user licences 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 misuse is difficult to detect, NERC carefully vets applications for data access and is prepared to take action against users caught in transgression of the user licence. |

### 3.7 Using the framework

The framework 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.

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 who have been or may become involved with them at other stages of their life cycle.

The framework is particularly relevant to the costs involved in preserving digital resources. Decisions taken throughout the life-cycle of the resource, especially at the design and creation stage, will have implications for preservation costs. The framework therefore underpins the cost model developed by Cimtech, which is discussed in chapter 5.