7. COMPLETING THE JIGSAW: MANAGING THE DIGITAL PRESERVATION PROCESS

7.1 Drawing the strands together

This report has attempted to outline the key issues in digital preservation and has examined in detail some of the processes involved and the interplay between them:

- The interaction of the stakeholders: their attitudes to digital preservation; the rights they expect to be upheld; the responsibilities expected of them and which they are prepared to fulfil.
- The life-cycle of the digital resource and the ways in which different stakeholders' interests influence stages of this life-cycle.
- The techniques for digital preservation, and the technological choices available.
- Tools for evaluating digital resources and selecting appropriate strategies for their long-term preservation.
- Identifying and estimating the costs of preservation.
- Ways in which digital resources may be lost or put at risk through neglect or damage. Methods of rescuing such resources.

Individual organisations will be able to pick out those issues and parts of the process that are of most relevance to their own interests and aims. Each is responsible for a different piece of the jigsaw. However, to achieve the ultimate goal—the long-term preservation of the cultural and intellectual heritage in digital form—we must recognise the interdependence between all the different stages in the creation, use and preservation of digital resources so that we can see the complete picture.

7.2 Managing a digital archive

Figure 3 represents a model for managing an archive facility. The central six functions, stretching from capture through to access and retrieval, are the core of the activity. The overall planning, reporting and administration are essential functions required to run the archive facility.

How the archive is managed and run operationally will have an effect on the whole preservation process. Efficient procedures during archive will reduce costs not only during capture, but also further back up the business chain of activities, possibly improving the efficiency of the creators themselves. Keeping up with and adapting to the changing technological environment will also influence costs and methods of retrieval and access.

The keys to the success of the archive are confidence and commitment. Increasing confidence in the security of the archive and the integrity of the items is essential to securing the commitment of stakeholders to it.

7.3 Managing The Digital Archive—constructing a ‘value chain’

This, in microcosm, is the essence of good practice in digital preservation. We can apply all the principles and procedures of managing a digital archive to the much larger picture of managing digital preservation as a whole.

Here, too, the key concepts are confidence and commitment. For the process to operate in an efficient and cost-effective way requires the commitment and cooperation of all the organisations and groups involved. By managing the complete process as a ‘value chain’, from capture to retrieval, improvements can be made in the way stakeholders communicate and work together, controlling costs, improving service levels, and raising quality levels.

7.4 A checklist for best practice

In order to achieve this ‘value chain’, organisations with an interest in digital preservation must complete their part of the jigsaw in as efficient a manner as possible. The following recommendations, proposed by AHDS, outline best practices to be adopted to ensure that data resources suit the purposes for which they are intended, and their content, appearance, and functionality may be cost effectively preserved over the longer term. They are not all directly relevant to all groups, but they are all essential to building the larger Digital Archive.

Data and collection design

Whether data resources are being created, accessioned physically into a collection, or made accessible to a defined user community, rigorous evaluation is essential to ensure that the resources serve the purpose for which they are intended at an affordable cost. Data creators, digital collection developers and data archivists should take note of best practice in the evaluation of potential data resources. That evaluation will, at a minimum, entail examination of the resources’:

> contents, scope and relevance to a defined user community or purpose;
> technical characteristics (structure) and how these influence the resources’ fitness for use and long-term maintenance;
> documentation (whether this is sufficient for the intended use and long-term maintenance);
> legal terms and conditions which attach to their 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.

Data creation

The way in which a data resource may be used, at what cost, and even whether it may be preserved, are influenced principally by decisions taken when creating or

including that resource in a collection. 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.

Data creators who are interested in ensuring the secondary use and preservation of their resources will adopt standards and best practices. While these reduce the cost involved in securing a data resource's long-term viability and subsequent use, they are not 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.

Data creators should consider the following categories of standards:

- those which ensure that data can be migrated with minimum content loss across platforms (standard file formats, and compression and encoding techniques);
- those which ensure that data can be migrated meaningfully between individuals and organisations (documentation standards, as devised by specialist communities and curatorial professions);
- those which ensure that data resources are comparable with other like resources (data value standards, as derived by specialist communities);
- those which ensure that data resources suit the purposes for which they are created (as derived by specialist communities.

Data storage and data structures

These should be chosen to support intended use and/or preservation scenarios.

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 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 users.

Many stakeholders will need to confront or implement remote data management strategies. These provide particular challenges, for example where remotely managed resources are to be integrated from a user's point of view with data and other information resources 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.

Data documentation

Data documentation is essential to the exchange of data resources between platforms and individuals. At a minimum, it should provide information about a resource's provenance, contents and structure, and about the terms and conditions attached to its subsequent management and use. It should be sufficiently detailed to support:

resource discovery (i.e. the location of a resource which is at least briefly described along with many other resources);
resource evaluation (i.e. the process by which users determine whether they require access to that resource);
resource ordering (i.e. the information which instructs a user about the terms and conditions attached to a resource and how to gain access to it);
resource use (i.e. the information which a user may need in order to access the resource's information content);
resource management (i.e. administrative information essential to a resource's management as part of a broader collection and including information about location, version control, etc.).

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 need to take note of standard documentation practices.

Those responsible for managing digital collections are usually best able to determine how data resources in that collection should be documented, while those responsible for creating resources within a collection are usually 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.

Preservation strategies

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.

Migration is not appropriate for the following kinds of resources. For these, technical preservation (except in the case of the second) or emulation may be preferred.

those in which the hardware platform makes an essential contribution to the resource's meaning and/or to the experience of its use (e.g. video games, game boys);
those where data are stored in undocumented proprietary formats (i.e. proprietary information systems which store data in binary formats);
those where data are stored in undocumented formats and bundled with access software which is also undocumented (e.g. commercial CD-ROM products).

Preservation practices (with regard exclusively to data migration)

Data preservation entails the creation and maintenance of archive copies of data files. Archive copies of a data resource are independent of any online representation and as such are distinguished from back-up copies of the same resource. Periodic and systematic back-up of online data resources is not in itself a sufficient preservation strategy.

Archive copies should be stored on industry standard digital tape or on other approved contemporary media as may arise.

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 latter.

Archive copies should be written with different software to protect data against corruption from malfunctioning or virus- or bug-ridden software.

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.

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 suppliers 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.

The integrity of data files should be checked periodically using checksum and similar procedures. Such procedures may be implemented automatically by the archive system according to parameters set by system operators

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 conduct a cost benefit analysis 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.

Data use

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 some way toward alleviating this fear and enhancing collection development activities. Investigation into the development and widespread deployment of such mechanisms is a priority.

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.

Figure 3: Managing a digital archive