5. ESTIMATING THE COSTS OF DIGITAL PRESERVATION

5.1 Isolating a 'preservation cost'

Deciding on the techniques for preservation is only part of the picture. A key question for everyone interested in digital preservation is—how much will it cost? One of the problems encountered in trying to answer this question is that all the aspects and tasks involved in digital collection management are closely interlinked, making it very difficult to identify those elements which relate solely to preservation. For the same reason, it is difficult to isolate that portion of a data centre's budget which is concerned only with preservation. Some costs do relate specifically to preservation, but this does not mean that those are the only costs incurred when preserving digital resources. For example, does the cost of creating an online catalogue or a printed catalogue count as a preservation cost?

Rather than attempting to isolate a global preservation cost, we should assume that there are some preservation costs associated with all the elements involved in the life-cycle of a digital resource.

5.2 A cost model

Cimtech has developed a cost model that can be used to establish and compare the costs of the preferred methods of preservation for each category of digital resource (Hendley, 1998). It is based on the integrated policy framework drawn up by AHDS, which was described in chapter 3. It defines seven key areas, based on the life-cycle of the digital resource, which must be addressed by any digital collection policy, of which preservation is one. Given that a successful preservation strategy depends upon good practice in all the other six areas in the framework, the cost model also needs to take into account the costs incurred in each of those other six areas. The seven areas are:

1. Data creation
2. Data selection and evaluation
3. Data management, including:
    - Data documentation
    - Data validation, including:
        - Data assessment
        - Data copying
        - Media refreshment
    - Data structure
    - Data storage
4. Resource disclosure
5. Data use
6. Data preservation
7. Rights management

The model is drawn from an analysis of the seven areas of this framework, in order to:

- define the tasks involved in implementing each of the key areas;
- identify the key costs associated with these tasks and indicate the factors which will reduce or increase costs;
- indicate those costs which are directly or indirectly related to preservation.

Figure 2. Cost model

	Area of framework	Elements to consider		Preservation costs
1	**Creation**	• Level of control • Overall costs • Best practice costs • Data centre costs	→	• Promoting best practice • Cleaning up digital resources
2	**Selection/evaluation**	• Collection policy • Technical/practical criteria for evaluation • Complexity	→	• Practical and technical evaluation
3	**Data management**	• Documentation • Validation (assessment, copying, refreshment) • Data structure conversion • Data storage • 'Data archaeology'	→	• Reading/editing/managing documentation • Validation (assessment, copying, refreshment) • Data structure conversion • Off-line archive storage • File recovery from media; cryptanalysis of files of unknown origin
4	**Resource disclosure**	• Online resource discovery agents • Gateways • Online catalogues	→	• Post-preservation disclosure costs
5	**Data use**	• Online access • Digital distribution media	→	• Post-preservation use costs
6	**Data preservation**	• Technology preservation		• Keep old hardware/ software • Third party service costs
		• Technology emulation		• Third party service costs
		• Digital information migration	→	• Change media (recording; archival storage) • Backward compatibility (set up; run; check) • Interoperability (test; set up; run; check) • Convert to standard forms (agree form; test; run; test delete)
7	**Data use/rights**	• Usage restrictions? • Depositor/user rights	→	• Clearance, management and authentication costs

5.3 Creation costs

Decisions made when a digital resource is created will have a significant impact on the options subsequently available for its future management, use and preservation. They will also significantly affect the cost of those options. Some organisations involved in preservation are able to exert considerable control over the creation process, while others have little or no influence.

Two organisations which have taken measures to control the life-cycle of the digital record—the Public Record Office (PRO) and the Natural Environment Research Council (NERC)—have already been described. Others are also in a position to influence the creators of digital resources. For example, the Arts and Humanities Data Service (AHDS) serves a specific community, which it has been able to target to explain the benefits of depositing digital resources with AHDS data centres. The Data Archive also specifies technical criteria and preferred data standards which depositors are encouraged to adopt. Increasingly, these centres will also be able to advise scholars and researchers in the humanities on the best practices for creating, documenting and depositing their digital resources.

Many other organisations, on the other hand, have no influence on their depositors. There will continue to be many cases where valuable digital resources will be discovered or deposited which have not been adequately managed or documented. In these cases collection managers will have to make difficult decisions about whether they can afford to take the necessary remedial action. Some of the issues and the techniques involved in ‘rescuing’ such data are discussed in the next chapter.

One of the biggest problems which data centres all face is in trying to ‘clean-up’ digital resources. The creators of a digital resource are best equipped to validate and document it. If they fail to do this, then the cost of ‘clean-up’ at a later stage, when most of the context will have been lost, is conservatively estimated to be ten times greater. In many cases it is impractical to attempt to clean-up digital resources retrospectively, resulting in rejection and loss of the data.

Clearly, the adoption of ‘best practice’ at the data gathering and creation stage can help simplify the task of managing and preserving the digital resource thereafter. And simplification will result in reduced costs.

Data centres, therefore, cannot ignore the creation stage—even if it is outside their direct control. They face two major cost areas relating to the creation stage.

- The cost of promoting good practice to depositors. This can include posting guidance notes on Web pages, running courses for defined user groups, educating funding bodies, and incorporating guidance notes in funding literature. All data centres should invest at least some resources in this preventative measure.
- The cost of correcting mistakes and examples of bad practice at the creation stage. The less spent on the first area, the more they will have to spend on the second.

On a per digital resource basis, the second costs will far outweigh the first. There is theoretically no limit to the amount of money that data centres could spend on

cleaning-up deposited digital resources. There are three practical ways of limiting these costs:

- Where the data centre has a fixed budget, allocate a maximum figure to 'clean-up' activities.
- Define a basic standard. Valuable resources which fall below that standard are brought up to that standard but not taken beyond it.
- Stipulate that digital resources which do not meet a minimum standard in areas such as documentation will be rejected.

Both the promotional (best practice) and the corrective (clean-up) costs relate directly and indirectly to preservation.

5.4 Selection and evaluation (acquisition) costs

The selection of digital resources will be based on the centre's collection policy. The costs associated with evaluating digital resources — assessing them against a series of technical and practical criteria — relate directly to preservation. How easily can the resources be managed, catalogued, accessed by end users and preserved by the data centre?

The level of cost involved will depend on the size and complexity of the digital resource and how well documented it is.

5.5 Data management costs

Data management covers all the tasks involved in managing a digital resource once it has been accepted into a collection.

The **documentation** supplied with a digital resource should describe its structure, contents, provenance and history. Centres must check the documentation, edit or add to it, if necessary, make it available to users and keep it up to date. There is a growing need to hold documentation in digital format, so if it is only available on paper there is an extra task involved in digitising it.

Even where the documentation provided is good, reading and studying it incurs a cost. Where it is poor, then clearly the costs increase dramatically as the centre has to test the digital resource and produce additional documentation. Managing the documentation and converting it to digital format can also prove very costly, particularly if the data centre holds a large number of digital resources and there is a significant volume of paper documentation for each resource.

The costs associated with reading, editing and managing the documentation for a digital resource are directly related to preservation. Without documentation it is extremely difficult to ascertain whether a resource can be preserved, or to determine the best strategy for its preservation.

Validation covers a number of procedures, which are together designed to ensure the integrity of the digital resource. They include:

- Assessment, to ensure that the resource is complete as documented, that it is functioning properly and operates on the specified hardware and software environments, and that it is consistent.
- Copying, to provide a back-up in the event of the media being destroyed or damaged
- Refreshing, to protect against the corruption that would result from any deterioration of the media.

This is an iterative process and so costs are repeated several times during the lifetime of a resource. The costs associated with assessing, copying and refreshing a digital resource all relate directly to preservation.

Data structure covers the way in which a digital resource has been formatted, compressed and encoded, together with any changes which the data centre may decide to make to these. The way in which the resource was originally created will play a large part in determining the level of cost involved in structuring it for future storage and access.

For example, if the resource was created on a proprietary application and deposited in a proprietary format not supported by the data centre, then significant costs would be involved in converting it to a standard format for long term storage and management. These costs clearly relate directly to preservation. However, the costs associated with conversion to a preferred format for delivery to users are outside the preservation remit.

Data storage options are influenced by the resources available, the volume of data to be stored and the way in which they will be used and preserved.

In determining costs we should distinguish between online or near-line storage set up to meet the access requirements of users and the management requirements of the centre, and off-line storage set up specifically to meet the preservation requirements of the centre.

None of the costs associated with the active management of digital resources relates directly to preservation (although a percentage of the costs, associated with migration., refreshing, etc., would have an indirect relation).

The costs associated with the off-line archive storage facility all relate directly to preservation.

‘**Data archaeology**’ may be required to process incomplete data or analyse files of unknown origin. This is discussed in detail in chapter 6.

5.6 Resource disclosure costs

Resource disclosure concerns the way in which information about a specific digital resource is made available to end users. This information will depend on the resource discovery tools that are implemented for the collection as whole. Examples include resource discovery agents, logically ordered gateways and online catalogues.

While the costs associated with these online resource discovery tools do not in most cases relate to preservation, once the object has moved from active life into an archive there are costs involved in discovering, extracting and preparing the object for use.

5.7 Data use costs

This area covers the costs associated with delivering the digital resources to end users. The costs depend upon the structure of the specific digital resource, how it is stored, and how the user needs to access it—which may involve conversion to different delivery formats. Various delivery mechanisms are possible, including online and publishing on CD-ROM. However, the costs associated with delivering the digital resources to end users are not related to preservation, again except where the delivery is from the digital archive.

5.8 Data preservation costs

Costs will clearly vary, depending on which preservation strategy is adopted.

Apart from copying and refreshing, which have already been dealt with as part of the validation exercise, the **technology preservation** strategy involves preserving the original application program, operating system software and computer hardware platform that which were all used to create or access the digital resource.

The costs involved to a data centre adopting this strategy would be of two types:

- After moving to a new hardware and software environment, the cost of keeping the old environment running for a short period of time while working on a migration strategy to cover those valuable digital resources that can only be accessed via applications that will only run on the old environment.
- If subsequently the data centre discovered any valuable old digital resources that could only be accessed via applications that would only run on the old environment, they would need to identify a third party still running the old hardware and software and pay them to load the application and the data, and to convert the data into a standard format which they would be able to preserve and migrate to the new environment.

Both types of cost relate directly to preservation.

Technology emulation involves the additional task of designing emulator programs to run on current and future computer platforms and programming them to mimic the behaviour of specific operating system software.

The cost would take the form of a fee to a third party to cover the use of facilities to emulate the required hardware and software environment, so that the centre could run the application and convert the digital resource to a standard format.

Again, this cost relates directly to preservation.

The four techniques within the **digital information migration** strategy can also be divided up into a series of tasks, each of which has a cost associated with it.

The costs associated with changing media are threefold:

- Formatting the digital resource and recording it on the new medium — for example, transferring from compact disk recordable (CD-R) to magnetic tape or from video disk to CD. Two copies are needed—one for active use and one for archive.
- Managing the new medium, including indexing, active and archival storage, and the provision of equipment for access.
- Making copies for distribution.

Of these, the recording costs and the archival storage costs relate directly to preservation. However, the costs of storing the active copy of the new medium and the costs of making copies for users are not related to preservation.

The costs associated with backward compatibility would be relatively low in the short to medium term, provided the applications remain backward compatible. If the collection manager decided to upgrade all files created on version X of the application to version X+1 then, provided it was feasible to write a macro that retrieved each file and saved a copy of each file in the new format, this process could be automated. The costs would include the following:

- setting up the macro and running it;
- checking a subset of the resulting files to ensure no corruption has taken place;
- deleting the previous version of the files (if the collection manager opts to do this).

All relate directly to preservation.

The third technique relies on interoperability between rival popular application programs. The more complex the digital resource the more difficult it is to interchange between two application programs without significant loss of data—and the higher the cost incurred.

As with backward compatibility, this process could be automated. The preservation costs include the following:

- testing the interchange on a range of representative documents;
- setting up the program and running it;
- checking a subset of the resulting files to ensure no corruption has occurred;

- deleting the previous version of the files if required.

The fourth and most popular technique is conversion to standard formats. The simpler the digital resource, the easier — and cheaper — it is to select a standard format and effect the conversion. In many cases the process would be identical to that just described for interoperability. After initial tests have been completed, digital images can be converted automatically from one compression algorithm to another and from one file format to another. Most documents created on Windows applications, for example, can be converted to PostScript files or to Adobe's PDF format.

The main costs which relate to preservation would include:

- agreeing on the preferred standard formats;
- testing the conversion for a specific category of resource;
- running the conversion as a batch process;
- testing a sample of converted resources;
- deleting the old versions if required;
- copying the resulting files.

5.9 Rights management costs

Rights management concerns all the processes involved in defining and upholding the rights of depositors and the users of the data centre. These include intellectual property rights and related legal issues such as data protection and confidentiality.

The rights vested in a resource may determine not only how the resource can be accessed and used but also how and whether it can legally be preserved by a third party. The costs associated with rights management can be substantial and may prove to be the highest cost area for the process of digital archiving.

5.10 The cost model in practice

The Cimtech cost model offers a schematic basis for examining digital preservation costs in more detail as those involved in the process explore the options available and the techniques involved. Practical experience will add flesh to the bones, so that in time more sophisticated models will evolve, which managers can begin to use when drawing up budgets and project plans.

The following example applies the model to a real-life organisation involved in the business of data collection and preservation.

The British Oceanographic Data Centre (BODC) was set up in 1989 as NERC's designated data centre for marine data to carry out the following functions:

- provide data management support for UK marine science;
- maintain and develop the UK's national oceanographic database;

- make oceanographic data available to UK academia, industry and government;
- develop innovative marine data products and digital atlases;
- collaborate on behalf of the UK in international exchange and management of oceanographic data.

Creation

Most of the data sets managed by BODC result from major experiments or surveys conducted at sea. Managing oceanographic data involves eight stages:

1. Pre-cruise planning
2. Pre-cruise calibration
3. Data collection at sea + log of activities
4. In situ calibration
5. Data processing, calibration and quality control, at sea and post cruise
6. Production of publishable data set (fully worked up, quality checked, labelled and documented)
7. Data as an end product
8. Application-independent database (available for secondary usage, to create other databases, products, etc.)

While BODC promotes the production of good data as an end in itself, the scientists involved in the experiments rarely have the time or resources to complete steps five and six. Once they have collected the data and carried out selective processing on subsets of interest to them, their first concern is the publication of scientific papers.

The result is that the data deposited with BODC are often incomplete and therefore unusable by others. The centre must spend time and resources processing the data—which would be unnecessary, had all the steps had been completed in the first place. BODC describes this frustrating and expensive task as ‘data archaeology’. The centre has tried to stipulate to depositors how data should be processed and documented, but scientists are overstretched and do not have the resources to do it. Often the task is delegated to junior staff and data can be lost.

Recently the centre has taken a different approach, providing a data management service to the cruises. BODC staff accompany the expedition in order to collect, process and document the data as they are created. The cost of providing this service is justified when compared to the overall cost of gathering the data, which often amounts to £10,000 per day. One recent project which has benefited from this approach is OMEX 1. The overall cost of the project was £10 million, while the data management cost £300,000—approximately 3 per cent of the total budget. Overall, staff have produced 600 data sets on CD-ROM, holding the data gathered on some 47 cruises.

Selection and evaluation

BODC is increasingly involved in the planning stage of new projects and, as just described, in managing the data gathering and management processes. Where the centre is heavily involved in the data creation stage there is less need to carry out selection and evaluation activities as the quality and integrity of the data have already been established before deposit.

Data management

As a general rule, BODC estimates that on a new project some 3–5 per cent of the budget goes on data management.

The centre always aims to publish the data for a project in an application-independent database so that it can be widely used and more easily preserved. Most of the data is alphanumeric, which can be loaded into databases. The centre is now starting to produce electronic documentation using Adobe Acrobat.

However, the data for half of the older experiments or cruises have never been worked up. They are archived, but cannot easily be published, so their value is diminished. This represents a significant data management problem.

It can take BODC five years to work up the data and produce published data sets for a major project. During that time the data sets are managed online using an Oracle relational database engine and made available with password protection to members of the team. All the data are backed up on recordable CDs and digital archive tape (DAT).

Resource disclosure

The bulk of the effort goes into structuring the data for each project, organising the data and providing finding aids for the data sets on CD-ROM. From operating principally as a data archive with some publishing activities, BODC is now primarily a data publisher with a back-up archive function.

Data use

The main method of data distribution is now via published CD-ROMs. A list of all the data sets and an ordering service are provided on the Web. Over the next five years the centre will be producing six CD-ROMs holding data sets gathered on 100 cruises.

Data preservation

The centre has adopted a digital information migration strategy. Staff devote a lot of effort to processing and managing the data, both at the creation stage and on receipt at the centre, to ensure they understand and can migrate them as required.

Part of the processing is designed to make the data application-independent. The data sets are loaded into standard relational tables so they can be imported into spreadsheets and a wide range of relational database engines.

Data sets stored on 4,000 magnetic tapes, built up when the centre started, have now all been transferred to recordable CDs. All new data sets are stored on recordable CDs and tape.

Rights management

The centre follows the policies outlined in the NERC Data Policy Handbook, so contract administration is usually a standardised process.

Overall preservation

In total eighteen BODC staff are engaged in the activities described above.

- six handle promotion of best practice and support to cruises (creation), and selection and rights management;
- eight handle data management and preservation;
- two handle resource disclosure;
- two handle data use.

Overall, six of the eighteen staff (one third) carry out activities which relate directly to preservation.