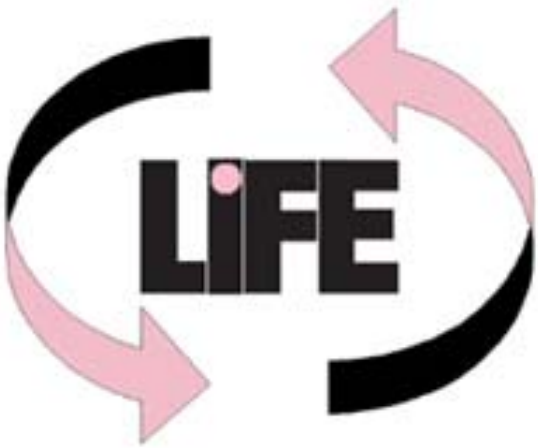
The LIFE Project

Bringing digital preservation to LIFE



Lifecycle Information for E-literature

Full Report from the LIFE project

A JISC funded joint venture project under 4/04 programme area (i)
Institutional Management support and Collaboration

Contents

Executive Summary

Introduction

The LIFE Project has developed a methodology to calculate the long-term costs and future requirements of the preservation of digital assets. LIFE has identified a number of strategic issues and common needs

The critical strategic issues are:

- The continuation of this research is crucially dependent on wider collaboration between Higher Education (HE) and Libraries and on cost-effective development of tools and methods.
- The time required for the realistic development of the next generation of these tools and methodologies is largely unknown and forms part of a wider collaborative responsibility within digital preservation.
- There exists a real opportunity to establish long-term partnerships between institutions to address common requirements. The challenge is to establish multidisciplinary Project teams and programmes to lead these developments;
- There exists a real opportunity to establish long-term partnerships between institutions and industry to develop this methodology and to establish new opportunities and to share knowledge and experience. The LIFE project could become an important vehicle for the development of these new opportunities

Method

The LIFE methodology is lifecycle based. The project was able to successfully use this approach to establish a cost to acquire and store digital content. The project also created a new Generic LIFE Preservation Model which leads to the project demonstrating that;

- The lifecycle approach to long-term custodianship and digital curation is feasible for any size digital repository and should be refined further.
- The Generic LIFE Preservation Model provides a solid foundation for the costing of preservation activity.

Cost

LIFE established that in the first year of a digital assets existence;

- The lifecycle cost for a hand-held e-monograph is £19
- The lifecycle cost for a hand-held serial is £19
- The lifecycle cost for a non hand-held e-monograph is £15
- The lifecycle cost for a non hand-held e-serial is £22
- The lifecycle cost for a new website is £21
- The lifecycle cost for an e-journal is £206

LIFE further predicts that in the tenth year of the same digital assets existence;

- The lifecycle cost for a hand-held e-monograph is £48
- The lifecycle cost for a hand-held serial is £14
- The lifecycle cost for a non hand-held e-monograph is £30
- The lifecycle cost for a non hand-held e-serial is £8
- The lifecycle cost for a new website is £6,800
- The lifecycle cost for an e-journal is £3,000

It is in this predictive work that further research is required. For example by year ten significant rises are measured for both web archiving and e-journals yet e-serials reduce. These figures come from a small sample of the collections and must be tested further to see if this is constant.

Preservation costs

The development of the Generic LIFE Preservation Model helped establish the cost to preserve digital assets within the lifecycle model but in isolation to other areas such as ingest and metadata. Further development of the model, integration with the broader lifecycle approach and refinement of its inputs using real data will be crucial in taking this forward.

Obsolescence watch

The project team conducted data mining and identified over 500,000 individual files made up of over 40 different file types. Large numbers of HTML and text files were encountered alongside more modest numbers of document and multimedia objects and smaller numbers of more unusual proprietary formats like GFF and ELEGANS. The majority of the collections examined were captured in the last two years with some going back as far as five years. None of the objects encountered were obsolete but the project considered some to be old and likely contenders for preservation action at some point in the near future. Continued vigilance to monitor digital collections will help to inform the frequency of necessary preservation action.

- LIFE encountered no obsolete formats in a five year old digital collection.

Collaborative understanding and tool development

Differences between institutional workflow proved challenging in the LIFE project, from acquisition and selection through to workflow and allocating costs. Most of these issues were overcome within the lifecycle model, however a conclusion from LIFE has to be that in order to be successful at collaborative work you must fully understand how your partner works. The greater the understanding of the differences and similarities, the higher the success ratio and the more realistic national standards and approaches become. LIFE strongly advocates this collaborative approach and would like to expand its experiences in this area to more accurately apply costs across a wider range of collections.

- The greater the collaboration between institutions, the greater the understanding of differences, the greater chance of success and standardisation

This collaborative approach extends to tool development; LIFE recommends support for collaborative tool development to be able to deal with a range of complex objects. Large scale reductions in cost can be expected with the correct tools. The high cost of ingest and metadata creation found in the project will continue if tools are not developed around normalisation at ingest and migration/emulation. For example ingest and metadata form around 60% of the total lifecycle cost for an e-monograph. This is an area where LIFE considers significant gains can be made.

- Collaborative tool development will significantly reduce the cost of ingest and metadata creation.

Executive summary conclusion

It is clear from the report that a price can be put against the lifecycle of digital collections. LIFE has made steady progress in one year to review existing models, choose a relevant methodology, customise this model and then test it against three diverse collections. LIFE established that it costs £19 to store and preserve an e-monograph which indicates that the model can be applied to digital collections. To be successful this work now needs to be continued in these summarised areas to test both the accuracy and relevance of this research within a wider collaborative HE/Library audience.

The following pages contain the full project documentation and Case Studies which led LIFE to these conclusions.

1. Introduction

This Report is a record of the LIFE Project. The Project has been run for one year and its aim is to deliver crucial information about the cost and management of digital material. This information should then in turn be able to be applied to any institution that has an interest in preserving and providing access to electronic collections.

The Project is a joint venture between The British Library and UCL Library Services. The Project is funded by JISC under programme area (i) as listed in paragraph 16 of the JISC 4/04 circular- Institutional Management Support and Collaboration and as such has set requirements and outcomes which must be met and the Project has done its best to do so. Where the Project has been unable to answer specific questions, strong recommendations have been made for future Project work to do so.

The outcomes of this Project are expected to be a practical set of guidelines and a framework within which costs can be applied to digital collections in order to answer the following questions;

- What is the long term cost of preserving digital material
- Who is going to do it
- What are the long term costs for a library in HE/FE to partner with another institution to carry out long term archiving
- What are the comparative long-term costs of a paper and digital copy of the same publication
- At what point will there be sufficient confidence in the stability and maturity of digital preservation to switch from paper for publications available in parallel formats
- What are the relative risks of digital versus paper archiving

The Project has attempted to answer these questions by using a developing lifecycle methodology and three diverse collections of digital content. The LIFE Project team chose UCL e-journals, BL Web Archiving and the BL VDEP digital collections to provide a strong challenge to the methodology as well as to help reach the key Project aim of attributing long term cost to digital collections.

The results from the Case Studies and the Project findings are both surprising and illuminating.

2. Background

The LIFE Project has set itself some challenging targets, just how much does it cost to acquire, ingest, store, access and preserve digital collections?

As you can imagine just finding a point to start was in itself a challenge and a comprehensive literature review formed the first part of the LIFE Project. Taking the Project team through a series of industries, from software development to construction the team searched for work previously done to form the basis for the start of the Project. Surprisingly it was somewhere closer to the Project partners' own business that the review ended and a Library model chosen.

From this decision the Project began to take shape and the chosen methodology was applied to a diverse range of digital collections. However while diverse in range there was little depth of information to any digital collection within either UCL or BL archives so data mining became a key Project requirement, the oldest collection chosen was actually only five years old but had little technical information. It did though set a key Project metric of having a preservation action every five years for this collection, a theme the team kept where possible across all the Case Studies.

The three Case Studies were chosen with the precise aim of challenging the methodology as robustly as possible. The model held up to scrutiny in all areas but with one key exception, preservation. This problem needed a specific solution for the lifecycle approach to succeed for the LIFE Project.
The British Library's new Digital Preservation Team and the DOM programme team developed a separate model for LIFE preservation specifically dealing with the complexity of creating costs for digital preservation through time. This Generic LIFE Preservation Model played a key role in the final Project outcomes.

So having added this final component the way was clear for the Project team to be able to produce a cost for a complete lifecycle for three digital collections. It meant that LIFE could provide real costs attributed to accurate data and produce figures for the majority for this Project.

This in turn has meant that the LIFE Project has taken the concept of cost within a digital environment and has delivered real price information which future Projects can use in order to build a clear cost picture in this complex and evolving area of collection management.

The following chapters outline in detail how this occurred. The literature review, the chosen methodology, the new Generic LIFE Preservation Model and the findings from the three Case Studies are to be found below. This Report then concludes with key themes, findings and areas for future Project work suggested.

3. The Literature review

In November 2005 James Watson completed a comprehensive review of existing lifecycle models and digital preservation in order to find a useable cost model that could be applied to the management of digital collections within a Library and HE/FE sector. This is a brief synopsis of the full 96 page review which is available on the LIFE website at www.ucl.ac.uk/ls/lifeproject/documentation/review.doc

This review introduced to the Project the concept of lifecycle costing (LCC) which is used within many industries as a cost management or product development tool. It is concerned with all areas of a product's lifecycle from inception to retirement. The review looked at LCC work within the construction industry, the product development industry and even the waste management industry to find an appropriate methodology.

However as it was within the Library sector LCC work that the greatest synergy was recorded and, given that the collections most likely to be considered for the Project were housed within Libraries, it made sense to review the work already done to cost the lifecycle of analogue Library collections to see if this could be directly transferable to the digital world.

This decision to follow a library trail led to a strong alignment with the work that was started in 1988 by Andy Stephens. In this work a formula for calculating the total cost of keeping an item in a Library throughout its lifecycle is introduced. No figures are attributed to the work at this point but the theory of a lifecycle approach is developed within the context of this work.

This work is significant as it is the first attempt found which takes a Library-based approach to the lifecycle management of assets, and although quite obviously developed for the paper world there is a strong correlation between the stages of analogue and digital asset management.

Stephens returns to this work in 1994 and allocates costs to specific parts of the National collection, namely serials and monographs. The findings indicate that costs vary for identical material dependent upon the procedures applied to the item within its lifecycle. For LIFE this sits well as the need for a formula, that can adequately cope with the many different varieties of electronic data and sources, had become the main point of focus.

This work was continued by Helen Shenton in 2002/03 where specific focus on the aspects of preservation costs throughout the lifecycle was included. This is a key extension and provides the first example of a lifecycle cost model with a consideration for preservation. It was decided at this point that a tool set in these terms would be the best fit and would be used by the LIFE Project.

4. The chosen methodology - a lifecycle approach

This section describes the LIFE Project's chosen model for digital materials. At first glance this may look like a challenging formula to use but in actual fact it is a powerful and relatively straightforward model to use to get a feel for the cost of managing any digital collection.

The accuracy of the output however is dependant on the sub layers and customisation added alongside the amount of real data that you have to put in to the calculator. The more data you collect or have, the more accurate the model becomes.

This Lifecycle Model is designed to fit to all digital library collections. The stages defined within are not compulsory, but rather provide a framework within which to work that will be applicable to most situations.

By allocating a cost to as many relevant sections as possible and by applying the Generic LIFE Preservation Model (see chapter 8) a total Lifecycle cost can be achieved.

4.1. Introduction

This section provides a generic breakdown of the different elements of a digital object's lifecycle. Calculating a summation of these elements over a specific time period will provide a complete lifecycle cost.

$$L_T = Aq + I_T + M_T + Ac_T + S_T + P_T$$

L is the complete lifecycle cost over time 0 to **T**.

It is intended to provide a broad enough scope to be usefully applicable to most digital collections while providing enough specific elements to allow a detailed lifecycle breakdown.

4.2. Lifecycle categories and elements

There are 6 main lifecycle categories that break down further into lifecycle elements. There may be special cases where additional elements are required for a specific implementation. Additional elements can be added at the end of the most appropriate category. Some elements will not be applicable, in which case they can be left blank (for example there is no effort or cost in the IPR element of the VDEP Case Study, as IPR is covered under Voluntary Legal Deposit Legislation).

Acquisition[1] (Aq)

- Selection (Aq1)
- IPR (Aq2)
- Licensing (Aq3)
- Ordering and invoicing (Aq4)
- Obtaining (Aq5)
- Check-in (Aq6)

Ingest (I)

- QA (I1)
- Deposit (I2)
- Holdings update (I3)

Metadata (M)

- Characterisation (M1)
- Descriptive (M2)
- Administrative and structural metadata (M3)

Access (Ac)

- Adding and maintenance of links. Reference linking (Ac1)
- User support (Ac2)
- Access mechanism (Ac3)

Storage (S)

- Bit-stream storage costs (S1)

Preservation (P)

- Technology watch (P1)
- Preservation tool cost (P2)
- Preservation metadata (P3)
- Preservation action (P4)
- Quality assurance (P5)

[1] Acquisition broadly relates to "Pre-ingest" in OAIS.

4.3. Explanation of stages

4.3.1. Acquisition (Aq)

Selection (Aq1)

Selection is the process of deciding which digital materials should be acquired.

Research review reference: See Schonfeld, King, Okerson, Fenton (2004) for description of this, (research review, page 14).

Case Study example: Selection process for UCL e-journals Case Study, selection of web site titles to gather in Web Archiving Case Study.

IPR (Aq2)

IPR covers the process of negotiating the rights to store, preserve and possibly provide access to the selected digital objects.

Research review reference: See CEDARS cost elements of digital preservation (sections 2 and 3), negotiating the rights to preserve the object and negotiating the right to provide access to the object.

Case Study example: Seeking permissions from web site owners in the Web Archiving Case Study.

Licensing (Aq3)

Licensing is related to IPR but specifically covers the process of negotiating the rights to access and to provide access to digital materials for a period of time.

Research review reference : See Schonfeld, King, Okerson, Fenton (2004), (research review, page 14).

Case Study example: Licensing negotiations for UCL e-journals Case Study.

Ordering and invoicing (Aq4)

This element covers the administrative and accounting processes of ordering and invoicing for digital objects, whether purchased or licensed.

Research review reference: Montgomery, Sparks (2000), (research review, page 8).

Case Study example: Ordering electronic journals in the UCL e-journals Case Study.

Obtaining (Aq5)

This is the process of acquiring the object from the source via whatever means (for example by post on handheld media, by email, by ftp).

Case Study example: Gathering web site instances in the Web Archiving Case Study.

Check-in (Aq6)

Check-in is a verification process to ensure that what was expected to be obtained actually arrives. It does not constitute a detailed Quality Assurance process that might verify that a specific digital object is what it purports to be (this can be found in the following Ingest category). Check-in is a less thorough process that might, for example, verify issues, titles or filenames that are expected.

Research review reference: Receipt and check in as in Schonfeld, King, Okerson, Fenton (2004), (research review, page 14).

Other

Other phases as applicable to specific collections.

4.3.2. Ingest (I)

QA (I1)

The Quality Assurance element represents the process of ensuring the obtained materials are of a sufficient level of or expected level of quality and applying fixes or re-acquiring the materials as appropriate. QA includes the process of checking the materials for viruses.

Research review reference: QA is contextualised by Harvard University Library (2002), (research review, page 27). Virus checking is encountered extensively in the research, but to cite a specific example see Jones, Beagrie (2001): Acquisition and appraisal, retention and review > Appraisal and selection > Procedures to prepare data and documentation for storage and preservation > Validation > Scanning for computer viruses (research review, page 74).

Case Study example: Verifying the quality of a gathered web site instance and providing manual fixes as appropriate in the Web Archiving Case Study.

Deposit (I2)

Deposit is the process of committing the digital entity to the repository, and any associated operations.

Research review reference: For a discussion of deposit within context see Hendley (1998), (research review, page 17).

Case Study example: Ingesting of the digital objects into the object management system in the VDEP Case Study.

Holdings update (I3)

This stage refers to the updating of holdings records on the systems of a library (catalogue, web pages, etc) when new content is accessioned.

Research review reference : This stage can be seen in context as a data collection instrument as described in King, Aerni, Brody, Herbison, Kohberger (2004), (research review, page 67).

Case Study example: This stage can be seen in the case histories as holdings update from the UCL e-journals Case Study, the Aleph procedures undertaken by the acquisitions staff in the VDEP Case Study, and the updating of the holdings database in the Web Archiving Case Study.

Other

Other phases as applicable to specific collections.

4.3.3. Metadata (M)

Characterisation (M1)

This process of characterising a digital object, analysing its properties and extracting metadata.

Case Study example: Characterisation is not currently performed in any of the Case Studies.

Descriptive (M2)

The application of descriptive metadata.

Research review reference: It is widely referenced in the research encountered in the Project. See Carroll, Hodge (1999) and Phillips (2005) for examples (research review).

Administrative and structural metadata (M3)

The application of administrative and structural metadata.

Research review reference: This stage can be found in James, Ruusalepp, Anderson, Pinfield (2003*), (research review, page 30).*

Case Study example: VDEP Case Study.

Preservation Metadata (M4)

This is the process of recording a variety of metadata for the purposes of preservation. Note that this should be considered the initial preservation metadata that is gathered and recorded around ingest time. Metadata updates recorded at the time when preservation activity is performed is covered below under P) Preservation.

Case Study example: Preservation metadata is not found in the Case Studies, but is discussed further in the Generic LIFE Preservation Model section.

Other

Other phases as applicable to specific collections.

4.3.4. Access (Ac)

Adding and maintenance of links. Reference linking (Ac1)

All activities involved with the setting up and maintenance of links to digital objects where necessary.

Research review reference: This activity is referenced in Schonfeld, King, Okerson, Fenton (2004), (research review, page 14).

Case Study example: Updating catalogue records in the Web Archiving Case Study.

Access mechanism (Ac2)

This represents the direct lifecycle costs associated with the mechanism to provide access to the digital materials.

Case Study example: The hosted access mechanism provided by Magus in the Web Archiving Case Study.

User support (Ac3)

Any activity covered under enquiry services, reference services and user support under correspondence (telephone, email, etc).

Research review reference: Montgomery, Sparks (2000), (research review, page 8). Most of the activity is under the data collection instrument Information Services.

Case Study example: Answering enquiries in the UCL e-journals Case Study

Other

Other phases as applicable to specific collections.

4.3.5. Storage (S)

Bit-stream storage costs (S1)

The bit stream storage costs.

A discussion of this element is found in Ashley (2000) and 39, Chapman (2003), (research review, page 33).

Case Study example: Internally provided storage is costed in the VDEP Case Study, and externally contracted storage is costed in the Web Archiving Case Study.

Other

Other phases as applicable to specific collections.

4.3.6. Preservation (P)

Preservation activities are not currently undertaken in any of the Case Studies but the issue is discussed in detail the Generic LIFE Preservation Model (chapter 8). The elements of P are summarised as:

Technology watch (P1)

Preservation tool cost (P2)

Preservation metadata (P3)

Preservation action (P4)

Quality assurance (P5)

Other

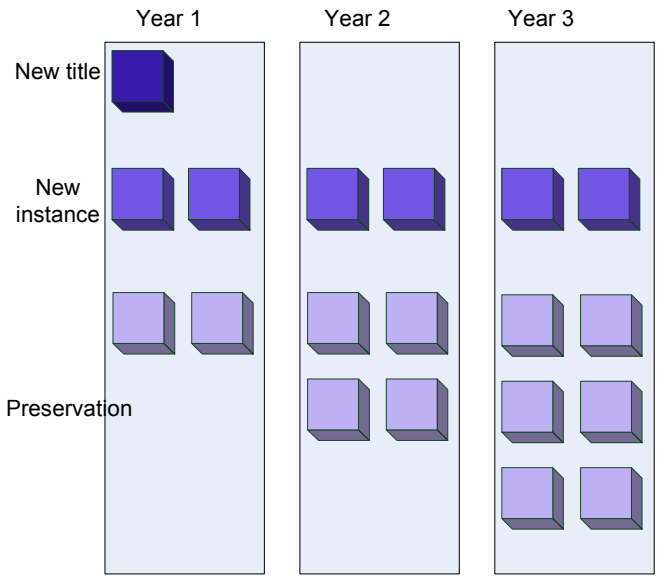
Other phases as applicable to specific collections.

4.4. Occurrence of Costs

Costs can occur at different stages of the lifecycle and can occur just once or a number of times at different frequencies. LIFE proposes the calculation of costs for a single title or entity over a specific time period. Examples of the range of occurrence of costs can be found in the Web Archiving Case Study, which includes:

- One off costs in the first year including Selection and IPR.
- Recurring costs for each instance of the title that is gathered, which includes a range of elements such as Obtaining, QA, and Deposit.
- Recurring annual costs for the preservation of each instance gathered.

Shown visually, there is a one off cost in the first year for selecting the new title, a cost for each new instance per year (in this example the title is gathered on a biannual basis), and a preservation cost for each instance gathered so far per year.


Year 1
Year 2
Year 3
New title
New instance
Preservation

5. VDEP Case Study

Author : Rory McLeod

This provides an overview and analysis of the lifecycle collection management for the Voluntarily Deposited Electronic Publications at the British Library. The lifecycle as analysed here aims to provide insight into the full life both current and projected of the VDEP material.

It was compiled in conjunction with the VDEP Manager, acquisitions staff, the e-media cataloguing team and the DOM programme staff; all of whom the LIFE Project thanks greatly for their help.

A full breakdown of the lifecycle cost over time for a variety of objects is given in section 5.5. An explanation how the lifecycle cost was applied is given below.

5.1. Introduction

This aims to analyse all aspects of the digital lifecycle of the VDEP collection. It also aims to highlights where efficiencies can be made and where additional processes could be implemented to aid preservation.

In some lifecycle costing work, "average" collection items are examined. If these objects are monographs, this would mean a monograph of average size, with serials this would mean a serial of average size with the average number of issues a year.

However, this approach has not always proved to be applicable in this context, due to the variation in the size, frequency and complexity of items within the collection. After extensive discussions with collection staff about what was possible and what was not, the LIFE Project elected to use Case Studies for certain parts of this.

- Where the costs do not vary according to file sizes and frequencies a generic per cost "average" item has been applied,
- Where the costs do vary per file size, type and frequency specific Case Studies have been supplied.

5.2. The collection

The VDEP collection has been arriving since 2000. At the time of writing, more than 230,000 separate objects have been deposited into the system. These separate items in turn make up 172,484 bibliographic records (or objects). A breakdown of file format types has been given in Table 5.

An exploration of these objects and items appears in more detail below. It is useful to consider the items that are being deposited as falling into four distinct categories. These are

1. Hand-held monographs (i.e. deposited on hand–held media such as CD-ROM)
2. Hand-held serials and issues,
3. Electronic serials (i.e. all others not deposited on hand-held media) and issues
4. Electronic monographs.

It is necessary to make these distinctions at this stage because of the difference in management processes and storage methods for each category. These differences are expanded below.

The items deposited in the collection are from a variety of sources. Although most are not, strictly speaking, academic journals, the file formats and modes of publication are similar enough for the collection to be an effective comparator for a collection of electronic journals. The profile of the file types in the collection, which is outlined below, is indicative of the similarity with e-journals.

Tools

The library management side of the collection is handled in Aleph. This covers adding new items, creating orders and skeletal records etc. The digital asset management side of the collection is handled by Ex Libris' digital asset management system DigiTool (version 2.3).

As always with digital collections the workflows and lifecycle stages outlined below are quite strongly governed by the system that is used.

5.3. The lifecycle

The following sections will analyse and cost each aspect of the VDEP lifecycle, categorised using the following formula:

$$L_T=Aq+I_T+M_T+Ac_T+S_T+P_T$$

L) is the complete lifecycle cost over T) the length of time in question

This approach is described in detail in section 4 of this Report.

5.3.1. Acquisition (Aq)

Selection (AQ1)

Selection is generally not a time-consuming stage for the VDEP material. The VDEP team's experience tells us that most things that have been deposited by a publisher have been accepted and processed by the team.

However, for context it is worth referring to the Guidelines for legal and voluntary deposit which state that the following non-print publications are excluded from Voluntary Deposit:

- Publications which substantially duplicate the content of a print publication from the same publisher, which has already been deposited;
- Publications published only for internal use within an organisation, including Intranet items and other materials clearly intended for a private audience, e.g. internal communications within a company, institution or other body;
- Publications already deposited under a publishing agreement;
- E-mail alert services principally pointing to web content, but containing only minimal content themselves;
- Online material clearly originating from outside the UK and Northern Ireland, the Republic of Ireland, Isle of Man and the Channel Islands, and which are primarily addressed to audiences outside these territories;
- Materials which are largely advertising a particular company's products, or
- are essentially an online form of mail order catalogue;
- E-cards, e.g. online birthday cards;
- Electronic games;
- Computer software; and
- Continuously updated publications such as 'dynamic' databases.

Since all material is voluntarily deposited and to a large extent sits within the boundaries set by Government, no staff time is expended on selecting or ordering the items.

Where used, check-in procedures are covered in the ingest module of the lifecycle.

IPR (Aq2)

N/A

Licensing (Aq3)

N/A

Ordering and Invoicing (Aq4)

N/A

Obtain (Aq5)

There is some physical processing of hand-held items. Specifically, these are the items that are sent to London to be stamped and given a physical shelf mark. However this type of work has at time of printing reduced to less than 5 CD-ROMs per month, and as such incurs no specific costs for processing.

Check-in (Aq6)

N/A

5.3.2. Ingest (I)

Ingest forms the first significant stage of the lifecycle process of the VDEP material. It has one-time cost elements and recurrent cost elements.

The ingest procedure is undertaken by the VDEP acquisitions team at Boston Spa plus the Manager in charge of production.
As well as the itemised costs for the four different types of material, which will be explained in the subsequent section, a Management cost should also be considered at this point to ensure all costs are considered.
The VDEP team Manager's time is spread across all areas of the VDEP process. It is therefore difficult to allocate this cost to a specific area and as such the cost has been spread over the total number of deposited objects up to 01/02/2006 (see Table 5).
20% time B Grade manager salary = £149.90 or £0.0013p per object. It was decided not to include this floating cost.

Hand-held monographs or first issues of hand-held serials

These are master records for new monographs and serials deposited on hand-held media

All the ingest costs below are one-off. This is due to the fact that it is the first and only time this monograph or serial will have been added to the collection. Subsequent issues for serials incur a lesser cost due to the master record created here.

QA (I1)

N/A

Deposit (I2)

Hand-held monographs are received in the Legal Deposit Office where they are stamped. The average cost based on time plus wages to do this equates to **£0.28** per item.

Holdings update (I3)

The first piece of work undertaken by the legal deposit team covers the receipt of the item. This ingest work is created on specialist library cataloguing software called Aleph and costs **£2.40** per new title (serial or monograph)based on local wage and time figures. This acquisitions work on Aleph includes all general serials administration work, such as linking to publishers, creating orders, creating the schedule for publication and creating a skeletal record. A large percentage of this holdings update work is in actual fact metadata, as such a percentage of this cost needs to be split and placed into the Metadata section (see table 1)

After the Aleph work, hand-held items are zipped and the zipped file is added to DigiTool. This process costs approximately **£0.96** per monograph and **£0.60** per serial and is a one-time cost. The files are *zipped to make them easily ingestible to DigiTool.

***Zipping (an explanation)**
Although the process of zipping makes it simple to ingest into DigiTool, it also makes it impossible to know what the object's constituent parts are, including file sizes, file formats, operating systems and the "look and feel" of the object. Obviously this is a concern within a digital preservation context and is not a policy that is likely to continue long-term. The decision to use zipped files was taken prior to a digital preservation team being formed and without any formal digital preservation discussion having taken place. It was quite simply and understandably a Project decision made to make the process more easily manageable while the VDEP Project acquired experience with the system.
The objects are then ingested into a folder on the DigiTool file store system. The process costs approximately **£0.57** per monograph item and **£0.50** per serial item.

At this stage the object has been effectively ingested into the object management system (DigiTool 2.3) and its arrival has been recorded on the library management system (Aleph).

The final process in the ingest procedure is a matching procedure between the Aleph acquisitions records, which are updated at the time of receipt, and the digital object record itself. On the DigiTool storage system this is known locally as the "match and merge" stage and is where the key metadata information is extracted from Aleph and used to populate the new metadata fields. The cost for this, based on staff time, is **£0.29** per monograph item and **£0.34** per serial item, again based on the time taken to process this request automatically.

Total cost per hand-held monograph (one-time): **£4.50**
Total cost per hand-held serial (one-time): **£4.12**

Hand-held serial issues

These are issues of serials which already have a master record created.

QA (I1)

N/A

Deposit (I2)

Hand-held serial issues follow the same pattern as new serials above and so incur a **£0.28p** stamping process.
However they will not require a new record to be created on Aleph as the master record has already been created for this serial title.

Holdings update (I3)

Each issue is recorded on Aleph and costs: **£1.44**. The cost is less than above because there are fewer procedures to carry out on the system.

As the items are hand-held, they are still zipped, and will consequently incur the same cost as above at £**0.60** per item.

The serial is then loaded into a folder on the DigiTool file store system costing **£0.50** per serial item.

The merging of the records is also still required. As this procedure is the same for first time issues of serials or for those ongoing, the unit cost is the same. However, the cost as incurred over the lifecycle of an object is recurrent, depending on how many issues of a particular journal arrive in a given space of time. The recurrent cost is **0.34p** per issue.

Total cost per issue of a hand-held serial (recurrent depending on the number of issues per year): **£3.16**

Electronic monographs

These are e-monographs that arrive via electronic delivery and not on hand-held media

QA (I1)

N/A

Deposit (I2)

Non hand-held VDEP serials fall into two categories: they can arrive by email or they are downloaded from the web; either way an item is downloaded and then added to Aleph; it is then ingested into DigiTool. The small cost of doing so is absorbed in I3.

Holdings update (I3)

The Aleph work to "arrive" an electronic monograph covering the same procedure as outlined above costs: **£2.88**

Electronic monographs differ from hand-held monographs as they do not have to be transferred from hand-held media. This brings the cost in as a £**0.24** download cost rather than a zipping charge. As with other monograph costs at this stage in the lifecycle, this expenditure is a one-time cost.

The merging procedure however is the same as outlined above, which incurs the same cost of: **£0.29**.

Total (one-time) cost per electronic monograph: **£3.41**

Electronic serials

These are serials that arrive via electronic delivery and not on hand-held media. This is the initial new title set-up cost.

QA (I1)

N/A

Deposit (I2)

Non hand-held VDEP serials fall into two categories: they can arrive by email or they are downloaded from the web. Either way, an item is downloaded and then added to Aleph; it is then ingested into DigiTool. The cost of downloading is absorbed in the process of I3

Holdings update (I3)

The Aleph processing of these items (new skeletal records, publication schedules, orders etc.) costs: **£2.88 per item.**

The object is downloaded to DigiTool. This cost is: **£0.24 per item**. This cost is the same cost as is incurred by checking-in an electronic monograph outlined above and is one-time in nature.

The necessary records are then merged at a cost of: **£0.34.**

As with other new title costs, these costs are one-time in nature.

The total, one-time, cost for new electronic serials first issues is: **£3.46.**

Electronic serial issues

These are serials that arrive via electronic delivery and not on hand-held media. This is the subsequent issue cost after a master record has been set-up for this title.

These costs are similar to the new title cost, but they do not require the creation of skeletal records, as outlined above, so the process is slightly less time-consuming.

QA (I1)

N/A

Deposit (I2) and Holdings update (I3)

The ongoing issues will each need processing which costs: **£0.38** for the Aleph work, **£0.24** to download and **£0.34** to merge.

The total (recurring) cost per issue of a normal electronic serial is: **£0.96**

The difference is that these costs recur as issues arrive over time, and so if issues arrive four times a month then this cost will be incurred each time.

Finding

As can clearly be seen the process of (I) and (M) through all 5 workflows is very closely linked for the VDEP material. In order to maintain clean numbers for the full lifecycle cost, a Project decision was made to split the total Ingest cost as

follows: 50% Ingest, 50% Metadata based on the allocation and recording of staff time.

Total Ingest cost:

VDEP Material	One-time cost	Recurring cost per issue	Total Ingest cost
Hand-held monographs (£)	4.50*	n/a	£2.25 one-off N/A
Hand-held serials (£)	4.12*	3.16*	£2.06 one-off £1.58 per issue
Electronic monographs (£)	3.41*	n/a	£1.70 one-off N/A
Electronic serials (£)	3.46*	0.96*	£1.73 one-off £0.48 per issue

Table 1: Ingest costs. *50% of this cost passes to Metadata

5.3.3. Metadata (M)

This section of the lifecycle covers all of the metadata applied to the VDEP serials and monographs. 50% of the costs in Table 1 will be allocated into this section to ensure clean lifecycle cost figures.

(M1) Characterisation
N/A. There are no automated analysis or Metadata extraction tools in use with the VDEP material.

(M2) Descriptive metadata and (M3) Administrative/Structural

Each new item, whether it is a serial or a monograph, is catalogued in Aleph to *full British National Bibliography* standard. This activity is undertaken by the dedicated e-media cataloguing team. All cataloguing costs are one-time costs.

Monographs
With monographs: an item is either emailed to the e-media cataloguing team, if it is an electronic publication, or it is sent through physically to be catalogued if it is hand-held. Hand-held items are installed onto local machines and catalogued from the screen; remote items are accessed and catalogued as usual.

For a new monograph, this activity costs: **£8.69**
There is little difference between hand-held and electronic items.

Serials
First issues of serials are catalogued in the same way as monographs. Serials are harder to catalogue than monographs and consequently take longer. It is not unusual for a new serial record to take in excess of 1 hour' However it should be pointed out that this cost is only ever needed once and all issues thereafter do not require this activity.

For a new serial, this activity costs: **£13.04**
There is little difference between hand-held and "normal" electronic items.

VDEP Material	One-time cost (M1)	50% cost from Ingest (I) goes into M2 and M3	Total Metadata cost
Hand-held monographs (£)	8.69	£2.25	£10.94
Hand-held serials (£)	13.04	£2.06	£15.10
Electronic monographs (£)	8.69	£1.71	£10.67
Electronic serials (£)	13.04	£1.73	£14.77
Hand-held issues	N/A	£3.16/2 = £1.58	£1.58 per issue
E-serial issues	N/A	£0.96/2 = £0.48	£0.48 per issue

Table 2: Metadata costs

Some technical, structural and preservation metadata is assigned by the staff ingesting the object. This is the **50%** cost mentioned in the last section and must be allocated here (Table 2 column 3).
This includes basics such as: the software used to process the item, file name, the file format and copyright restrictions. There is some system-assigned technical metadata, including the file size and a checksum value (effectively a system ID).

It is worth noting that for electronic records, and consequently VDEP material, all metadata fields are under review. Full technical, structural and preservation metadata are currently under development through The British Library Core Metadata Group, which will deliver in 2006. For LIFE this means that the allocated cataloguing costs are at present a very manual and labour intensive process which is expected to decrease rapidly over time.

Also the lack of technical metadata reflects the limitations of the systems in use. DigiTool 2.3 has no support for any significant technical or administrative metadata. However, if a different system were to be used, it could be possible to install a template for the input of more technical, structural and preservation metadata.

Example: Certain processes could be automated and put onto a template. Specifically, technical metadata can be extracted from ingested files using JHOVE (or equivalent) and pre-defined preservation metadata can be input to a saved METS template (or equivalent), using appropriate preset authority files, including templates for inputting values for the PREMIS Object and Event entities.
http://www.loc.gov/standards/mets/
http://www.oclc.org

Recommendation: For the relatively simple and homogeneous objects that are ingested into the VDEP archive, it may not be a very expensive exercise to input preservation metadata (P) to an agreed standard as part of the existing workflow. This process could be carried out at ingest point with more complex technical and structural metadata possibly being referred to more specialist staff.

For instance, if an object had many more files associated with it than was usual, or if any of those associated files were unusual file formats, then this part of the ingest procedure could be referred on to a specialist team.
Some technical metadata is recorded already, so it would not be unknown for the staff undertaking the acquisitions process to record this information (for example, on hand-held items the required operating systems and hardware are added as metadata), much of the metadata could be recorded automatically. Given that – excluding hand-held items - in the collection there are only 20 file formats, of which 3 **(HTML, PDF, txt) account for over 85%** of all files, presets for files could easily be used.

Finding: it will be far more challenging and expensive to apply preservation metadata to hand-held resources due to their complex file structure and relationships. This must be explored in future work and has fallen out of the scope of this initial Project.

VDEP Material	One-time cost	Recurring cost per issue
Hand-held monographs (£)	Too early to say	Too early to say
Hand-held serials (£)	Too early to say	Too early to say
Electronic monographs (£)	Too early to say	Too early to say
Electronic serials (£)	Too early to say	Too early to say

Table 3: Technical, structural and preservation metadata costs
The costs here are flagged as Too early to say

5.3.4. Access (Ac)

Adding and maintenance of links (Ac1)

Access mechanism (Ac2)

User support (Ac3)

N/A
At the moment, there is no access to the DigiTool OPAC. Consequently there is no specific cost associated with access to the item that is not covered elsewhere in this Report. However, other mechanisms, such as storage and descriptive cataloguing, arguably do include access costs within them. Furthermore as the descriptive catalogue records are publicly available for the Integrated Library

System (ILS) OPAC, these items can be requested via the usual method. It is also quite straightforward to link these catalogue records to the digital object itself, thereby providing access.

If off-site access to digital objects were to be offered, this would create more maintenance of password files or maintenance of IP address registers and so on, creating greater complexity and higher access cost.

Finding: It is outside the scope of this Project to try to predict cost and timeframes for access to this collection. Indeed access when possible will require changes in both legislation and process to accommodate this.
There is no access to VDEP information due to its voluntary nature, so no costs for access can be attributed.

VDEP Material	One-time cost	Recurring cost per issue
Hand-held monographs (£)	n/a	n/a
Hand-held serials (£)	n/a	n/a
Electronic monographs (£)	n/a	n/a
Electronic serials (£)	n/a	n/a

Table 4: Access costs

5.3.5. Storage (S)

Bit-stream storage (S1)

The digital objects are, at the time of writing (January 2006), being stored in the DOM system, the British Library's Digital Object Management system. Before that, between 2003 and 2005, the items were stored on a standard filestore system. Both storage systems are interfaced with a DigiTool 2.3 front end. The DOM system is designed to be a digital archiving system which is fully equipped to preserve the items that come into it. The standard filestore system, which used about 2TB of storage capacity, was used between 2003 and 2005 as a temporary store. This system was a more straightforward "filestore" rather than a preservation system; within the confines of this Project, costs can be obtained per amount of file space. In actual fact, it is a good comparison between two very different systems – both in a start-up phase and both with a reasonable amount of real data to be extracted.

As the time line goes on, and the VDEP material is fully housed in the DOM system, year-to-year maintenance and Preservation costs can be extracted and a robust and predictable set of data added to this start point. All of these recurrent costs should be input as the cost of running an OAIS-compliant archival repository. For the purposes of this Project, a start has been made and any obvious costs associated with the running of this system have been used here.

All bit-stream storage prices should be viewed within context, and particularly appropriately to this Project, within the broader lifecycle of the management of a digital object. Bit-stream preservation costs are not without issues however; it is certainly a cost element of the lifecycle and gives a valuable metric which is directly comparable to the storage cost for a physical volume. The problems with using these figures for assessing long-term preservation are well known, and Ashley (1999) and Chapman (2003) provide excellent introductions to them. In short it is a mistake to class safe bit-stream storage costs as full-scale preservation costs. However, the comparison between Chapman 2003 costs and LIFE 2006 costs do give a useful comparison in straight bit-stream preservation costs.

As was explored by Chapman, the costs for bit-stream storage are comparable to those for storage of physical items and will vary in a similar way. For example: a small, environmentally-controlled storage facility in an area with high property prices will cost more than a large non-environmentally controlled facility in a low cost area. Similarly, a small OAIS-compliant system will cost more than a large standard file store. This should be kept in mind when comparing any costs, and is certainly applicable to LIFE due to the UK locations of the systems in question. In fact, to quote from the 2003, "emerging models for digital preservation reaffirm the fact that not all storage environments are equal". Chapman's study outlines a figure of:

Chapman 2003
$60/gb for <100 gb
$32/gb for 101-1000 gb and
$15/gb for >1000 gb for the OCLC digital archive

What we do not know is the security or redundancy built into this cost, so whether this is a straight comparison to the DOM system is difficult to ascertain. We also do not know if these costs represent full start-up costs or whether they are just representative of Hard Drive storage over time only.
In fact Chapman's conclusion gives us a good introduction to the LIFE Project's storage analysis.

The 2003 Report concludes: "Thus managed storage costs are not fixed, but arrived at collection-by-collection by judicious decision making. The choice of repository, the scope of service, the repository pricing model and owner's decisions regarding formats, number of items, number of versions and number of collections to deposit are all potential variables…."

LIFE 2006
And so, a major objective of the LIFE Project was to obtain costs for bit-stream storage to compare with these findings.

After extraction of the information surrounding the BL's file store system where the VDEP material is stored, a cost came out at **£80.85** per gigabyte for the first year that the file store was being used. Assuming that the storage system would

last for 5 years (and indications are that it will do this easily) this figure drops to **£16.17** per gab - this includes software upgrades and licenses over this 5-year time span and is the figure used as an average for the lifecycle cost.

It is important to note the large difference in start-up costs to the actual working figure which averages itself over time. As systems develop and grow, better metrics will become available.

This is a rough cut figure, but will serve as an approximation of the file store "rent". It is well worth noting that this system was purchased at a high specification and with a dedicated proprietary system. A comparative set of figures has been produced using what is known of the new DOM storage system's open configuration, and it serves as a good comparison as to just how much this can affect long-term storage costs. The equivalent DOM costs start at **£27.44** per gb for the first year and then drop to **£5.48** by the fifth year, which again will be the figure used within this work.

Table 5: Comparison of costs VDEP vs. DOM

System components	**VDEP File store 02/05**	**DOM start-up system***
Storage/server	138894	190761
Application	22095	30067
Back-up system	12000	4345
Installation	6584	0
Maintenance	25027 (over 3 years)	0
Compiler (VDEP)	578	n/a
DSE 1 (DOM)	n/a	25000
DSE 2 (DOM)	n/a	17400
Dark Archive (DOM)	n/a	98179
Support	0	95517
Additional software	1811	66784
Total Cost	**£206,989**	**£533,981**
Total gb available	2560	19456
Cost per gb 1st year	£80.85	£27.44
Cost per gb 5th year	£16.17	£5.48
Cost per gb 10th year	£8.08	£2.74
Cost per gb 20th year	£4.04p	£1.37p
Storage utilised (Jan 06)	1741	n/a
Number of files (Jan 06)	231,733	n/a
Average cost per file	0.089p	n/a
Average file size storage utilised /number of files)	0.0075695gb or 7.75mb	Using VDEP experience we can assume 8mb

These figures are a useful comparison for any long-term predictions of costs, as they are compiled for two completely different systems but are populated with real data. You might also say that they are examples of both best and worst case scenarios in the digital repository sense. VDEP's filestore was built with a state-of-the-art mentality which utilised the highest quality drives where the emerging

long-term strategy for DOM has taken a much more pragmatic approach to storage.

Future costs however should where possible be applied through time by applying industry trends in cost versus storage, for example by applying a "Mores law" type of approach to these trends. Research for this exemplar found that it was difficult to extract meaningful industry trends freely, so projected costs are based on the real costs experienced by both the DOM and VDEP Projects.

Some useful work was, however, done for the Digital Preservation Coalition under the Technology Watch Report "The large-scale archival storage of digital objects" at http://www.dpconline.org/graphics/reports/index.html#lgescale. This gives us an indication of what we might find. Overall, the findings indicated a decline in hard drive cost of between 30-40% alongside a doubling of space every 12 months. If this was to be proven correct, then we could project the following figures for the next 5 years to gauge a start-up cost for a repository based on the DOM approach.

Table 6: Projected cost to start up a repository system.

Year	Year 1	Year 2	Year 3	Year 4	Year 5
Cost	£533,981	£347,087	£225,606	£146,643	£95,318
gb	19456gb	38912gb	77824gb	155648gb	311296gb
Cost per gb	£27.44	£8.91	£2.89	£0.94p	£0.31p

However, it is highly speculative to project these equations forward as there are concerns around technological blocks and the fact that this rate of cost reduction and increase in storage could be unsustainable and will bottom out at some point.

Object file size

There are additional problems for storage related to the "size" of a digital object. If an "average" lifecycle cost for an object is to be extracted, then the "average" size of an object needs to be obtained. However, this is not an easy metric to obtain and is in fact not a helpful one in predicting long-term lifecycle costs.

Monographs are more straightforward than serials, but, in the VDEP collection, they vary so much in size that the results would not be to meaningful.

The average "size" of a serial is also problematic. New accessions will have fewer stored issues than older ones and therefore the figures will be skewed; furthermore the frequency of issues will also affect this figure.

Finding: For both serials and monographs there is no way of considering compound objects.

To circumvent this problem, the Project will allocate storage costs by applying the costs to specific items (titles) within the collection. These costs can then be scaled up or down according to the type of projection required. Both the VDEP cost for 1, 5, 10 and 20 years and the DOM equivalent have been used to populate the tables in this section.

(S1) Bit-stream storage continued (Project examples)
In each case the 1-, 5-, 10- or 20-year cost for bit stream storage has been multiplied by the cumulative space taken up by each example.

Hand-held Monographs

- *Instructor's CD for engineering economic analysis*

This monograph consists of one CD-ROM which is just over 17mb.

System	year 1	year 5	year 10	year 20	Total
VDEP	£1.34	£0.27	£0.13	£0.06	£9.00
DOM	£0.45	£0.09	£0.04	£0.02	£3.00
Size	17mb	17mb	17mb	17mb	17mb

1 Gigabyte =1024 Megabytes

- *Lancashire 1851 census*

This monograph consists of 35 CD-ROMs which average about 600mb each, giving a total of about 21 gb.

System	year 1	year 5	year 10	year 20	Total
VDEP	£1697.85	£339.57	£169.78	£84.89	£11460.45
DOM	£576.24	£115.24	£57.62	£28.81	£3889.55
Size	21gb	21gb	21gb	21gb	21gb

1 Gigabyte =1024 Megabytes

- *The number crew: measures, shapes and spaces*

This monograph consists of one CD-ROM, which is 587mb.

System	year 1	year 5	year 10	year 20	Total
VDEP	£46.34	£9.26	£4.63	£2.31	£312.70
DOM	£15.73	£3.15	£1.57	£0.79	£106.20
Size	587mb	587mb	587mb	587mb	587mb

1 Gigabyte =1024 Megabytes

- *AGI source book for geographic information and systems*

This monograph consists of one CD-ROM, which is 2276kb (2.233 mb).

System	year 1	year 5	year 10	year 20	cumulative
VDEP	£0.18	£0.04	£0.02	£0.01	£1.25
DOM	£0.06	£0.01	£0.01	£0.003	£0.42
Size	2.233mb	2.233mb	2.233mb	2.233mb	2.233mb

1 Gigabyte =1024 Megabytes, 1 Megabyte = 1024 Kbytes

Hand-held Serials

- *OAG data*

This hand-held serial is monthly with an additional monthly supplement. Each issue averages at 100mb in size. Because, effectively, 2 issues per month are arriving. This equates to 2.34 gb per year.

System	year 1	year 5	year 10	year 20	cumulative
VDEP	£189.19	£189.19	£189.19	£189.19	£3783.80
DOM	£64.20	£64.20	£64.20	£64.20	£1284.00
Size	2.34gb	11.7gb	23.4gb	46.8gb	46.8gb

1 Gigabyte =1024 Megabytes

- *Belfast working papers in Language and Linguistics*

This hand-held serial is on CD-ROM and the average size per issue is 1380kb. This serial is published intermittently (on the evidence of the two deposited copies, once every two years). This equates to 690kb per year.

System	year 1	year 5	year 10	year 20	cumulative
VDEP	£0.05	£0.05	£0.05	£0.05	£1.00
DOM	£0.01	£0.01	£0.01	£0.01	£0.20
Size	690kb	3.36mb	6.74mb	13.47mb	13.47mb

1 Gigabyte =1024 Megabytes, 1 Megabyte = 1024 Kbytes

Electronic Monographs

As with hand-held items, there is no easy way of interrogating the system to reveal the average sizes of non hand-held publications. This factor is additionally complicated by the fact that there is also no way of extracting the differences between monographs and serials. Consequently, the Project will adopt the same approach as above.

- *Measurement requirements and methods for optical fibre polarisation controllers*. This is a monograph submitted in 2004 which is 1.66mb in size.

System	year 1	year 5	year 10	year 20	cumulative
VDEP	£0.13	£0.026	£0.013	£0.0065	£0.87
DOM	£0.04	£0.008	£0.004	£0.002	£0.27
Size	1.66mb	1.66mb	1.66mb	1.66mb	1.66mb

1 Gigabyte =1024 Megabytes, 1 Megabyte = 1024 Kbytes

- *European incumbents get down to core business: strategies for success*. This monograph is 150mb in size.

System	year 1	year 5	year 10	year 20	cumulative
VDEP	£11.84	£2.37	£1.18	£0.59	£79.90
DOM	£4.01	£0.80	£0.40	£0.20	£27.05
Size	150mb	150mb	150mb	150mb	150mb

1 Gigabyte =1024 Megabytes, 1 Megabyte = 1024 Kbytes

Electronic Serials

- *Circulation*

This is a quarterly journal, with each issue being 0.256mb. As there are 4 issues per year this is a recurrent cost. First year= 1mb

System	year 1	year 5	year 10	year 20	cumulative
VDEP	£0.08	£0.08	£0.08	£0.08	£1.60
DOM	£0.03	£0.03	£0.03	£0.03	£0.60
Size	1mb	5mb	10mb	20mb	20mb

1 Gigabyte =1024 Megabytes, 1 Megabyte = 1024 Kbytes

- *E-Law*

This is a monthly journal, with the average size of each issue being 398mb. As there are twelve issues a year, this is a recurrent cost. First year 4.66 gb

System	year 1	year 5	year 10	year 20	cumulative
VDEP	£376.76	£376.76	£376.52	£376.52	£7530.40
DOM	£127.87	£127.68	£127.68	£127.68	£2557.40
Size	4.66gb	23.3gb	46.6gb	93.2gb	93.2gb

1 Gigabyte =1024 Megabytes, 1 Megabyte = 1024 Kbytes

Findings

If depreciation is used on hard drive storage over time, then the cost of the increasing serial (assuming each issue is identical) storage is identical through time.

If the metric to obtain the storage cost is the percent of the total storage capacity that an item takes up, then this exercise shows that storage costs vary considerably according to the size of the item. A range of £1.25 to £11,500 indicates how difficult the selection process is to cost.

5.3.6. Preservation (P)

This section of the lifecycle concentrates on the costs of the preservation of this collection. Preservation is an important part in the lifecycle of a digital object and there is a need to apply any cost metrics that can be extracted from the collections at this stage. In an environment consisting mainly of modern mainstream formats, which are stored in an OAIS-compliant repository, where the operating system is not an issue, risk remains in the file formats to be found there. It is in the spirit of all lifecycle costing exercises to apply current information to try and give information about future trends, and the LIFE approach follows this track.

If a lifecycle cost is applied to a digital collection, then there is a need for a time metric to be incorporated into the analysis. This time metric should ideally come from historical data which can then be extended through time. This will give an average deterioration rate for a specific collection which, assuming there is no paradigm change in computing, can be used to predict the amount of future preservation activity.

However, given that the VDEP collection is not old enough to have had large-scale digital preservation activities applied to it, there is no retrospective data to

be applied. On this basis, an assessment of the collection "as is" has been undertaken, with the time metric supplied by the fact that the collection has been in existence for 5 years and has experienced no preservation assessment or preservation activity in this time. So this forced interaction at the 5-year point has become our time metric for many areas of the Project.

Migration example

An example follows below which concentrates on migration at 5 years as a digital preservation strategy. This is not an endorsement of migration over any other digital preservation strategy, but merely an attempt to provide a cost of preservation for this specific collection.

Example of VDEP files as of 01/02/06:

File Format (extension)	% of collection	No of files	Obsolete	Migrated
att	0.18	424	No	Untested
bmp	<1	110	No	YES*
csv	1.11	2583	No	Untested
db	0.03	92	No	Untested
doc	2.25	5236	No	Untested
gif	1.16	2699	No	YES*
html	32.72	75851	No	Untested
jpg	2.70	6267	No	Untested
law	<1	82	No	Untested
li	<1	1	No	Untested
msg	<1	37	No	Untested
mso	<1	15	No	Untested
pcx	0.23	553	No but issues	YES*
pdf	10.52	24384	No	Untested
png	<1	12	No	Untested
ppt	<1	2	No	Untested
rtf	0.39	926	No	Untested
txt	42.65	98838	No	Untested
vcf	<1	21	No	Untested
xls	1.66	3863	No	Untested
xml	1.09	2532	No	Untested
zip	3.12	7236	No	Untested
Total	**99.89%**	**231773**	**None**	

Table 5: VDEP file formats
*Migration MIME types selected

Using VDEP as an example and in the truest of senses, **£0.00** is the cost for preserving this collection**:** VDEP is a voluntary scheme for testing the requirements of The British Library's Legal deposit obligations in future years. This does not of course mean that there is no cost just that the cost is delayed until digital preservation of this material comes with the DOM system and the

new Digital Preservation team strategy later in 2006. The material on deposit during this voluntary period may not even be considered for DOM, as a large part of the digital preservation process is hoped to be automated at ingest NOT after.

However, it is not possible to come up with a full lifecycle cost without considering some kind of digital preservation. Although The British Library at this point (January 06) is still developing a full strategy, LIFE is a project environment and as such it is possible to do some tests to estimate what it might cost to preserve the VDEP collection long-term.

VDEP file format findings

On examining the file formats in the collection, it is apparent that none has become obsolete. However, some are certainly ageing and are becoming less supported by software, a good example of this is GIF files. GIF files have been the most widely-used web graphic format. However, they are now less used mainly due to their proprietary nature. They are also an old file format created in1989 and have to a large extent been replaced by the use of the PNG (Portable Network Graphic, see: http://www.w3.org/Graphics/PNG/). They have also been used in digital preservation Case Studies as an "obsolete" file format, such as Rosenthal et al 2005: http://www.dlib.org/dlib/january05/rosenthal/01rosenthal.html. GIF files account for about 1.16% of the collection with a total of 2669 files, so they were a logical format to migrate within this exercise. This finding of format vulnerability triggered a preservation response and was treated as follows: analysis of the files was conducted to establish two more formats to go alongside GIF as a the file formats that were the oldest and under some sort of threat.

The results showed us that the most likely file formats that are waning in popularity and use within VDEP seem to be GIF, BMP, and PCX. These accounted for only a small amount of the overall collection, but nevertheless were the files most in need of attention. They accounted for the following:

- BMP files (110 items, 0.045% of the collection)
- PCX files (553 items, 0.261%)
- GIF files (2669 items 1.16%)

Out of interest over 75% was covered by txt and HTML files. However neither is deemed to be under immediate threat. It would be untrue, however, to say that this means that there are no concerns for the complexity surrounding HTML and its relationships with other programs and languages is a major issue. However, again, it was felt that this should come under future studies due to these complexities and the limited time available.

So using PNG as our preferred common format to migrate to, tests were run to better understand the time and cost associated with migration.

Migration exercise tools

2050 GIF, PCX and BMP files were extracted from the VDEP database. As these were all image formats, Photoshop 6.0 was used as a tool for migration. These files were then interrogated for authenticity using JHOVE. All three formats were then migrated using a script written using actions within the Photoshop 6.0 software into the more modern PNG format. The resulting timeframes were

GIF 1610 files
PCX 247 files
BMP 193

All migrated successfully in 960 seconds (16 minutes).

The error rate was 0

Each file migrated from a threatened format to PNG in 0.468 seconds.

Findings

If pre-emptive preservation is considered in this context, it becomes clear that migration is a feasible and appropriate option. All file formats mentioned above have a more modern stable format that is easy to migrate to (PNG) and this migration can be effected with relative ease with normal desktop software. The process to migrate these files is:

1. Locate file in system (flag as being in danger)
2. Open in migration software (Photoshop 6.0 in this case)
3. Create batch processing script and check for success
4. Run batch and migrate to common format
5. Re-ingest to the system recording the preservation action in the preservation metadata. (In this case date modified added to filename)

For this exercise, Photoshop was a valid choice and gives us a real cost. The cost of Photoshop at 2006 prices is £457.08 inc. VAT.

During our development work for this Project, we decided that 5-year intervals is the minimum timeframe we would leave a collection without some sort of digital preservation activity (even if it was just a check), so all costs for this exercise use this metric.

Tool
Therefore, total cost of preservation activity for this exercise, over five years is £457.08/2050 = **£0.22 per file** for the tool.

Set-up
The time taken to set up and run the batch process worked out to being 1 hour of set-up time (C Grade wage) to set up. This equalled £17.39/2050 = **£0.008 per file.**

To process
Rather than attempt to cost a computer running per hour, the operator-per-hour price has again been used to gain a per file cost – in this case £17.39 per hour, which is £0.29 per minute x 16 minutes, which gives a cost of £4.64/2050 or **£0.002 per file** to batch process.

To validate
10% (205) of the files were visually checked for accuracy against the original file. This was a quick check of file size, properties and visual representation. Tool development in this area should speed things up considerably but the process of visual validation in the absence of such was two hours. Visual validation therefore was £17.39 x 2 = £34.78/205 **= £0.17p per file**.

The per file total for this 2050-file exercise to this point then becomes

Tool £0.22
Setup £0.008
Batch £0.002
Validate £0.00
Visual £0.17
Total for Migration = £0.40p per file
LIFE could not further validate the success of this migration and future work is required to fully test the model.

Findings

The tool cost needs to remain at £0.22p as Photoshop is not able to access any of the HTML files within VDEP. However, if all the 231,773 items had for argument's sake been image files, the cost per file would drop to £0.002p so it is easy to see how a tool which can be applied over any format would vastly reduce the overall cost.

It is likely that a visual check would not be required and would be part of this developed tool, so the 0.17p per file allocated here would not be applicable. However for this exercise it is all that we have, so we must allocate the full £0.40p per file cost.

This £0.40p must now be split between base cost of migration and the base cost of testing a preservation action. This split is **£0.22p** for migration and **£0.18** for testing. This is then allocated to our Generic LIFE Preservation Model.

Bringing this together

To extrapolate further, and to give us our final lifecycle cost, we need to apply a cost across the entire 172,484 entities within. This could also be done across items (231,773), but for this exercise complete entities have been chosen. Unfortunately we must employ a little leap of faith at this point in tool development. Photoshop will of course not handle many of the formats in the

VDEP collection and no tool has yet been developed to handle all the varieties within.

Finding, Some major tool development work is required in this area. The more flexibility this tool has, the more the overall cost will drop.

However for this migration exercise we are going to assume it would take the same set-up and verification times for all VDEP files. In reality, for this to be viable, 0.468 seconds has to be close to the maximum time allowable or the exercise becomes unsustainable. So this time is maintained for the cost but flagged as a maximum.

On top of this we do have to allocate the development costs of tools and our other model costs (see Generic LIFE Preservation Model, chapter 8, for explanations)

The following information was extracted from the model to cost the digital preservation of this collection at 1,5,10 and 20 year intervals.

File Format (extension)	Format Complexity (FCX)	Estimated number of objects in one year	1 year cost	5 year cost	10 year cost	20 year cost
Txt	0.1	98838	£5,181	£11,370	£18,177	£29,283
Html	0.3	75851	£9,101	£17,821	£26,229	£36,311
Pdf	0.8	24384	£15,810	£28,117	£37,439	£39,716
Zip	0.4	7236	£8,122	£15,500	£21,773	£26,355
Jpg	0.2	6267	£4,951	£10,462	£15,882	£22,759
Doc	0.8	5236	£14,247	£25,165	£32,925	£32,547
Xls	0.8	3863	£14,135	£24,954	£32,602	£32,033
Gif	0.2	2699	£4,850	£10,272	£15,591	£22,296
Csv	0.1	2583	£3,316	£7,850	£12,795	£20,736
Xml	0.3	2532	£6,376	£12,677	£18,363	£23,818
Rtf	0.8	926	£13,895	£24,501	£31,909	£30,933
Pcx	0.2	553	£4,790	£10,157	£15,416	£22,018
Att	0.2	424	£4,786	£10,150	£15,406	£22,001
Bmp	0.2	110	£4,777	£10,134	£15,380	£21,961
Db	1.0	92	£16,844	£29,119	£37,157	£33,508
Law	0.8	82	£13,826	£24,371	£31,710	£30,617
Msg	0.3	37	£6,283	£12,502	£18,095	£23,393
Vcf	0.3	21	£6,282	£12,501	£18,093	£23,390
Mso	0.8	15	£13,820	£24,361	£31,695	£30,592
Png	0.2	12	£4,774	£10,128	£15,372	£21,948
Ppt	0.8	2	£13,819	£24,359	£31,692	£30,587
Li	1.0	1	£16,834	£29,102	£37,131	£33,467
	Total :		**£206,819**	**£385,573**	**£530,835**	**£610,268**
	Average cost per year :		**£206,819**	**£77,115**	**£53,083**	**£30,513**
	Average cost per entity :		**£0.89**	**£0.33**	**£0.23**	**£0.13**

Analysis of costs

Having used the figures from the migration exercise and the Generic LIFE Preservation Model we can extrapolate as follows.

The total cost to preserve this collection as it stands in the first year is £206,819. If this is then split across all entities (172,484) (and the assumption is that if it is not able to be retrieved as a whole then Preservation has failed) the cost is £0.89p per entity.

At subsequent trigger points the migration cost for (P) is;

Trigger point	Item
Year 1	£0.89p
Year 5	£0.33p
Year 10	£0.23p
Year 20	£0.13p

5.4. Conclusion

This Report had made some assumptions but is based mainly on fact. VDEP is however only one collection. The Lifecycle Model has been evaluated, modified and applied to this collection successfully. Given this fact, further work is strongly recommended tying in other collections and partners across the UK to evaluate how these costs stand-up to scrutiny.

Key Finding

Significantly, a five-year-old collection of published content has no technically obsolete files. Some files are ageing and deprecated by the web community, and these are described in this Report. However, these files are still supported and usable. This key finding means that while the current electronic information environment persists file formats, on today's evidence, are unproblematic. Although this is, perhaps, controversial it is difficult to think of a digital preservation project where "real" obsolete content was encountered. In fact our experience tells us that in the UK, many people are trying very hard to find obsolete formats (see Rusbridge 2006). That does not mean formats are safe, far from it, but it does give us some reassurance that the digital fear gripping the land is a little exaggerated. Although it is clear that all computer files will be obsolete one day, it is also clear that due to breadth of utilisation and the greater awareness of the problem, files are perhaps not as fleeting as they once were.

The digital preservation strategy modelled (i.e. manual batch migration) will not be applicable in all situations, but the exercise does demonstrate that digital preservation has a cost and based on informed and studied best estimates, is affordable with certain caveats.

It is also well documented that digital collections are mutable and do not fare well with benign neglect. Therefore to counter this, very close attention should be paid to:

- file formats in the existing collection (technology watch)
- proportions of file formats coming in
- the metadata applied to the file formats that are being ingested (standards and collaboration)
- Scheduled audits (i.e. risk assessments) of the collection "as is" to assess danger of obsolescence against agreed criteria to inform digital preservation priorities
- technical staff need to be available to develop tools (training)
- As much work as possible is released from the technical digital (standards and collaboration) preservation staff and is worked into ingest and acquisition procedures. Where possible, this work should be automated and scaled-up to reduce cost.

In conclusion, all digital collections should be treated with their lifecycle in mind. The more this approach is taken, the better the information retrieved. The better that this lifecycle information is managed, the easier digital preservation will be.

5.5. Full lifecycle costs applied to specific digital examples below.

In order to get a better understanding of what the total lifecycle cost is, and so we can apply this methodology to specific collection examples, the lifecycle costs are now put into practice. These examples will then be combined to give an overall average for the four different VDEP types; hand-held monographs/serials and non hand-held monographs/serials. Average cost can then be extracted for all four.

Total lifecycle costs for VDEP examples

- Hand-held Monographs

Title: *Instructors CD for engineering economic analysis*
1 CD-ROM 17mb in size

Element	Yr1	Yr2	Yr3	Yr4	Yr5 Preservation action	TOTAL
Aq	0	0	0	0	0	0
I	2.25	0	0	0	4.50	6.75
M	10.94	0	0	0	2.25	13.19
Ac	0	0	0	0	0	0
S	0.45	0.36	0.27	0.18	0.09	1.35

P	0.89	0	0	0	0.33	1.22
TOTAL for Years 1-5	14.53	0.36	0.27	0.18	7.17	22.51

Element	Yr6	Yr7	Yr8	Yr9	Yr10 Preservation action	TOTAL
Aq	0	0	0	0	0	0
I	0	0	0	0	4.50	4.50 (11.25)
M	0	0	0	0	2.25	2.25 (15.44)
Ac	0	0	0	0	0	0
S	0.08	0.07	0.06	0.05	0.04	0.3 (1.65)
P	0	0	0	0	0.23	0.23 (1.45)
TOTAL for Years 1-10						29.79

The total cost to store and preserve this e-mono for one year is £14.53
The total to store and preserve this e-mono over five years is £22.51
The total cost to store and preserve this e-mono over ten years is £29.79

- Hand-held Monographs

Title: *The number crew: measures, shapes and spaces*
1 CD-ROM 587mb in size

Element	Yr1	Yr2	Yr3	Yr4	Yr5 Preservation action	TOTAL
Aq	0	0	0	0	0	0
I	2.25	0	0	0	4.50	6.75
M	10.94	0	0	0	2.25	13.19
Ac	0	0	0	0	0	0
S	15.73	12.60	9.45	6.30	3.15	47.23
P	0.89	0	0	0	0.33	1.22
TOTAL for Years 1-5	29.81	12.60	9.45	6.30	10.23	68.39

Element	Yr6	Yr7	Yr8	Yr9	Yr10 Preservation action	TOTAL
Aq	0	0	0	0	0	0
I	0	0	0	0	4.50	4.50 (11.25)
M	0	0	0	0	2.25	2.25

						(15.44)
Ac	0	0	0	0	0	0
S	2.81	2.50	2.19	1.88	1.57	10.95 (58.18)
P	0	0	0	0	0.23	0.23 (1.45)
TOTAL for Years 1-10	2.81	2.50	2.19	1.88	8.55	17.93 (86.32)

The total cost to store and preserve this e-mono for one year is £29.81
The total to store and preserve this e-mono over five years is £68.39
The total cost to store and preserve this e-mono over ten years is £86.32

- Hand-held Monographs

Title: *AGI source book for geographic information and systems*
1 CD-ROM 2.2mb in size

Element	Yr1	Yr2	Yr3	Yr4	Yr5 Preservation action	TOTAL
Aq	0	0	0	0	0	0
I	2.25	0	0	0	4.50	6.75
M	10.94	0	0	0	2.25	13.19
Ac	0	0	0	0	0	0
S	0.06	0.05	0.03	0.02	0.01	0.17
P	0.89	0	0	0	0.33	1.22
TOTAL for Years 1-5	14.14	0.05	0.03	0.02	7.09	21.33

Element	Yr6	Yr7	Yr8	Yr9	Yr10 Preservation action	TOTAL
Aq	0	0	0	0	0	0
I	0	0	0	0	4.50	4.50 (11.25)
M	0	0	0	0	2.25	2.25 (15.44)
Ac	0	0	0	0	0	0
S	0.01	0.01	0.01	0.01	0.01	0.05 (0.22)
P	0	0	0	0	0.23	0.23 (1.45)
TOTAL for Years 1-10	0.01	0.01	0.01	0.01	6.99	7.03 (28.36)

The total cost to store and preserve this e-mono for one year is £14.14
The total to store and preserve this e-mono over five years is £21.33
The total cost to store and preserve this e-mono over ten years is £28.36

These three project examples now give us an average cost for hand-held e-monographs over time. These averages will be used to populate the final findings and executive summaries.

Example	Yr1	Yr5	Yr10
Instructors CD-Rom	14.53	22.51	29.79
The numbers crew	29.81	68.39	86.32
AGI Geographic	14.14	21.33	28.26
Average cost for e-monographs	**19.49**	**37.41**	**48.12**

- Hand-held serials

Title: *OAG data*
Issue per year: 24
Average storage size per issue: 100mb

Element	Yr1 new record	Yr1 issues x 23	Yr 2 x24	Yr 3 x24	Yr4 x24	Yr5 Preservation action	TOTAL at 5 years
Aq	0	0	0	0	0	0	0
I	2.06	1.58 (36.34)	1.58 (37.92)	1.58 (37.92)	1.58 (37.92)	3.16x120 (379.20)	531.36
M	15.10	1.58 (36.34)	1.58 (37.92)	1.58 (37.92)	1.58 (37.92)	1.58x120 189.60	354.80
Ac	0	0	0	0	0	0	
S	2.79	2.79 (64.17)	2.79 (66.96)	2.79 (66.96)	2.79 (66.96)	2.79x120 (334.80)	602.64
P	0.89	0.89 (20.47)	0	0	0	0.33x120 (39.60)	60.98
TOTAL for Years 1-5	20.84	6.84 x 23= 157.32	5.95 x 24= 142.80	5.95 x 24= 142.80	5.95 x 24= 142.80	7.86 x 120 =943.20	**1550**

Element	Yr6 X24	Yr7 X24	Yr8 X24	Yr9 X24	Yr10 Preservation action	TOTAL at 10 years
Aq	0	0	0	0	0	0
I	1.58 (37.92)	1.58 (37.92)	1.58 (37.92)	1.58 (37.92)	3.16x240 (758.40)	1441.44 (910.08)

M	1.58 (37.92)	1.58 (37.92)	1.58 (37.92)	1.58 (37.92)	1.58x240 (379.20)	885.68 (530.88)
Ac	0	0	0	0	0	0
S	2.79	2.79	2.79	2.79	2.79	1284.00
P	0	0	0	0	0.23x240 (55.20)	116.18
TOTAL for Years 1-10	5.95 x 24= 142.80	5.95 x 24= 142.80	5.95 x 24= 142.80	5.95 x 24= 142.80	7.76 x 240 = 1862.40	**3727.30**

The total cost to store and preserve this e-serial for one year is £20.84
The total to store and preserve this e-serial over five years is (£1550/120) £12.91
The total cost to store and preserve this e-serial over ten years is (£3727/240) £15.83

- Hand-held serials

Title: *Belfast working papers in language and linguistics*
Issue per year: 0.50
Average storage size per issue: 1380kb = 690kb per year
Number of files per issue: 1

Element	Yr1 new record	Yr1 no issue	Yr 2	Yr 3 x1	Yr4	Yr5 Preservation action	TOTAL at 5 years
Aq	0	0	0	0	0	0	0
I	2.06	0	0	1.58	0	3.16 x3 9.48	13.12
M	15.10	0	0	1.58	0	1.58 x3= 4.74	21.42
Ac	0	0	0	0	0	0	0
S	0.01	0.01	0.01	0.01	0.01	0.01	0.05
P	0.89	0	0	0	0	0.33 x3 = 0.99	1.88
TOTAL for Years 1-5	18.06	0.01	0.01	3.17	0.01	5.08x3 = 15.24	**36.50**

Element	Yr6	Yr7 x1	Yr8	Yr9 x1	Yr10 Preservation action	TOTAL at 10 years
Aq	0	0	0	0	0	0
I	0	1.58	0	1.58	3.16 x5	18.96

					15.80	(32.08)
M	0	1.58	0	1.58	1.58 x5 7.90	11.06 (32.48)
Ac	0	0	0	0	0	0
S	0.01	0.01	0.01	0.01	0.01	0.05 (0.10)
P	0	0	0	0	0.23 x 5 = 1.15	1.15 (3.03)
TOTAL for Years 1-10	0	3.17	0	3.17	24.85	**67.69**

The total cost to store and preserve this e-serial for one year is £18.06
The total to store and preserve this e-serial over five years is (£36.50/3) £12.16
The total cost to store and preserve this e-serial over ten years is (£67.69/5) £13.53

These two project examples now give us an average cost for a hand-held e-serial over time. These averages will be used to populate the final findings and executive summaries.

Example	Yr1	Yr5	Yr10
OAG data	20.84	12.91	15.83
Belfast working papers	18.06	12.16	13.53
Average cost for Hand-held e-serial	**19.45**	**12.53**	**14.68**

- Electronic Monographs

Title: *Measurement requirements and methods for optical fibre polarisation controllers*
Submitted 2004 1.66mb in size

Element	Yr1	Yr2	Yr3	Yr4	Yr5 Preservation action	TOTAL
Aq	0	0	0	0	0	0
I	1.70	0	0	0	3.40	5.10
M	10.40	0	0	0	1.70	12.10
Ac	0	0	0	0	0	0
S	0.04	0.03	0.02	0.01	0.008	0.108
P	0.89	0	0	0	0.33	1.22

TOTAL for Years 1-5	**13.03**	0.04	0.04	0.04	5.47	**18.59**

Element	Yr6	Yr7	Yr8	Yr9	Yr10 Preservation action	TOTAL
Aq	0	0	0	0	0	0
I	0	0	0	0	3.40	3.40 (8.50)
M	0	0	0	0	1.70	1.70 (13.80)
Ac	0	0	0	0	0	0
S	0.008	0.007	0.006	0.005	0.004	0.03 (0.14)
P	0	0	0	0	0.23	0.23 (1.45)
TOTAL for Years 1-10	0.04	0.04	0.04	0.04	5.37	5.53 **23.89**

The total cost to store and preserve this e-mono for one year is £13.03
The total to store and preserve this e-mono over five years is £18.59
The total cost to store and preserve this e-mono over ten years is £23.89

- Electronic Monographs

Title: *European incumbents get down to core business*
Submitted 2005 150mb in size

Element	Yr1	Yr2	Yr3	Yr4	Yr5 Preservation action	TOTAL
Aq	0	0	0	0	0	0
I	1.70	0	0	0	3.40	5.10
M	10.40	0	0	0	1.70	12.10
Ac	0	0	0	0	0	0
S	4.01	3.05	2.30	1.55	0.80	11.71
P	0.89	0	0	0	0.33	1.22
TOTAL for Years 1-5	**17.00**	3.05	2.30	1.55	6.23	**30.13**

Element	Yr6	Yr7	Yr8	Yr9	Yr10 Preservation action	TOTAL
Aq	0	0	0	0	0	0
I	0	0	0	0	3.40	3.40 (8.50)
M	0	0	0	0	1.70	1.70 (13.80)
Ac	0	0	0	0	0	0
S	0.68	0.56	0.44	0.32	0.20	2.20 (13.91)
P	0	0	0	0	0.23	0.23 (1.45)
TOTAL for Years 1-10	0.68	0.56	0.44	0.32	5.53	**7.53 (37.66)**

The total cost to store and preserve this e-mono for one year is £17.00
The total to store and preserve this e-mono over five years is £30.13
The total cost to store and preserve this e-mono over ten years is £37.66

These two project examples now give us an average cost for a normal e-mono over time. These averages will be used to populate the final findings and executive summaries.

Example	Yr1	Yr5	Yr10
Optical fibre	13.03	18.59	23.89
European Incumbents	17.00	30.13	37.66
Average cost for Hand-held e-serial	**15.01**	**24.36**	**30.77**

- Electronic serials

Title: *Circulation*
Issue per year: 4
Average storage size per issue: 0.256mb = 1mb per year

Element	Yr1 new record	Yr1 issue x3	Yr 2 X4	Yr 3 X4	Yr4 X4	Yr5 Preservation action	TOTAL at 5 years
Aq	0	0	0	0	0	0	0
I	1.73	0.48 x 3	0.48 x 4	0.48 x 4 =1.92	0.48 x 4 =1.92	0.96 x 20	28.13

		(1.44)	=1.92			=19.20	
M	14.77	0.48 x 3 (1.44)	0.48 x 4 =1.92	0.48 x 4 =1.92	0.48 x 4 =1.92	0.48 x 20 =9.60	31.57
Ac	0	0	0	0	0	0	0
S	0.03	0.03	0.03	0.03	0.03	0.03	0.15
P	0.89	0.89 x 3 (2.67)	0	0	0	0.33 x 20 =6.60	10.16
TOTAL for Years 1-5	17.42	1.88 (5.58)	3.87	3.87	3.87	35.43	**70.01**

Element	Yr6 X24	Yr7 X24	Yr8 X24	Yr9 X24	Yr10 Preservation action	TOTAL at 10 years
Aq	0	0	0	0	0	0
I	0.48 x 4 =1.92	0.48 x 4 =1.92	0.48 x 4 =1.92	0.48 x 4 =1.92	0.96 x 40 =38.40	46.08 (74.21)
M	0.48 x 4 =1.92	0.48 x 4 =1.92	0.48 x 4 =1.92	0.48 x 4 =1.92	0.48 x 20 =19.20	26.88 (58.45)
Ac	0	0	0	0	0	0
S	0.03	0.03	0.03	0.03	0.03	0.15 (0.30)
P	0	0	0	0	0.23 x 40 =9.20	9.20 (19.36)
TOTAL for Years 1-10	3.87	3.87	3.87	3.87	66.83	**82.31 (152.32)**

The total cost to store and preserve this e-serial for one year is £17.42
The total to store and preserve this e-serial over five years is(£70/20) £3.50
The total cost to store and preserve this e-serial over ten years is (£152/40) £3.81

- Electronic serials

Title: *E-Law*
Issue per year: 12
Average storage size per issue: 398mb (4.66gb first year)

Element	Yr1 new record	Yr1 issue x11	Yr 2 X12	Yr 3 X12	Yr4 X12	Yr5 Preservation action	TOTAL at 5 years
Aq	0	0	0	0	0	0	0
I	1.73	0.48 x	0.48 x	0.48 x	0.48 x	0.96 x 60	81.89

		11 5.28	12 =5.76	12 =5.76	12 =5.76	=57.60	
M	14.77	0.48 x11 5.28	0.48 x 12 =5.76	0.48 x 12 =5.76	0.48 x 12 =5.76	0.48 x 60 =28.80	66.13
Ac	0	0	0	0	0	0	0
S	10.64	10.64 x 11= 117.04	10.64 x 12= 127.68	10.64 x 12= 127.68	10.64 x 12=127. 68	10.64 x 12=127. 68	638.40
P	0.89	0.89 x 11 =9.79	0	0	0	0.33 x 60 =19.80	30.48
TOTAL for Years 1-5	28.03	137.39	139.20	139.20	139.20	233.88	**816.90**

Element	Yr6 X12	Yr7 X12	Yr8 X12	Yr9 X12	Yr10 Preservation action	TOTAL at 10 years
Aq	0	0	0	0	0	0
I	0.48 x 12 =5.76	0.48 x 12 =5.76	0.48 x 12 =5.76	0.48 x 12 =5.76	0.96 x 120 =115.20	138.24 (220.13)
M	0.48 x 12 =5.76	0.48 x 12 =5.76	0.48 x 12 =5.76	0.48 x 12 =5.76	0.48 x 120 =57.60	80.64 (146.77)
Ac	0	0	0	0	0	0
S	10.64 x 12= 127.68	10.64 x 12= 127.68	10.64 x 12= 127.68	10.64 x 12= 127.68	10.64 x 12= 127.68	638.40 (1276.80)
P	0	0	0	0	0.23 x 120 =27.60	27.60 (58.08)
TOTAL for Years 1-10	139.20	139.20	139.20	139.20	328.08	**884.88 (1702)**

The total cost to store and preserve this e-serial for one year is £28.03
The total to store and preserve this e-serial over five years is(£816/60) £13.61
The total cost to store and preserve this e-serial over ten years is (£1702/120) £14.18.

These two project examples now give us an average cost for a normal e-serial over time. These averages will be used to populate the final findings and executive summaries.

Example	Yr1	Yr5	Yr10

Circulation	17.42	3.50	3.81
E-Law	28.03	13.61	14.18
Average cost for Hand-held e-serial	**22.72**	**8.55**	**8.99**

This concludes the section on project specific costs, the importance of the numbers from this section and their comparison with the other case studies will be highlighted in both key findings and executive summary.

5.6. Notes on preservation

1) Arguably, all digital files will become obsolete with time. This has had a number of effects on the VDEP collection. PDF documents, HTML files and Word documents have been collected in the above collection for five years. The files that were submitted earliest in the life of the collection will be early version of the files. However, due to the system being used, it is impossible to extract information about which version of a given file format is in evidence.

1.1) Html, the universal language of the web, accounts for 33% of the VDEP archive. The application of strictly-coded HTML is notoriously unreliable when dealing with the open web. To this end, if preservation was the objective, it would be provident to take the versions of HTML and convert them into XHTML (i.e. HTML rendered as valid XML, see: http://www.w3.org/TR/xhtml1/); this would not only achieve normalization of the files to XML, itself a best guess future proofing activity, but would ensure cross-browser accessibility. This migration can be manually achieved relatively easily, with someone of the necessary skill set using standard desktop software. However, given that there are 70, 000 HTML files in the archive, this would cost an enormous amount of money. It would be more cost-effective to build this process into ingest procedures. It would also be more cost-effective to record what version of the file formats is being ingested and whether this marks up to a valid DTD.

2) Retrospective/batch normalization and migration
The above exercise is based an assessment of the likely preservation issues surrounding aging file formats in a real 5-year-old archive. One would find a very different picture of the costs of digital preservation if one took this collection and performed a risk assessment on the basis that, for preservation, file formats should be: "platform-independent, vendor-independent, non-proprietary, stable, open and well supported" (taken from recommended file formats for the FCLA digital archive, available at: http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf). If one were to migrate all files from a low confidence level file format to a high confidence level file format - which is debatably akin to retrospective normalization - one would be looking at the migration of about 40% of the file formats in the collection. In this collection, this means about 86, 000 files. This would take considerable staff resource: again, if this process was undertaken on ingest, there would be

considerable cost saving and better scope for preservation. If normalization was affected on ingest, the retrospective digital preservation activity would still be likely to be unnecessary.

2.1) Normalization can be considered a "first preservation activity" (see Shelton, 2003: http://liber.library.uu.nl/publish/articles/000033/english.html) the cost stage of which is under the ingest stage of the lifecycle.)

3) Further analysis of the digital preservation costing exercise tells us that, if effective policies are in place early in the lifecycle, preservation costs can be cut dramatically.

4) The costing exercise in this Report is based on the assumption that a linear deterioration rate will continue through time. This is, of course, not an assurance, but given that the archive has been in existence for 5 years, it is not an unfair assumption. Obviously this pattern will not continue indefinitely. When there is a paradigm shift in the way that digital information is accessed, all costs will be unpredictable

5) Migration here is taken to mean what is described in OAIS as transformation: "A Digital Migration where there is some change in the Content Information or PDI bits while attempting to preserve the full information content." P 5-5 OAIS blue book

6. Web Archiving Case Study

Author : Paul Wheatley

6.1. Introduction

This Case Study will consider the costs of the British Library's Web Archiving activities. Currently the BL is leading collaboration with five other institutions as part of the UK Web Archiving Consortium (UKWAC). The other members are:

- JISC
- National Library of Wales
- National Library of Scotland
- The National Archives
- Wellcome Trust

The aim of the UKWAC is to develop a test bed for archiving UK web sites, and selectively collect and preserve a cross section of culturally significant web sites. This is very much a learning process and the BL intends to scale up its Web Archiving operations in advance of likely legal deposit legislation in this area. For the purposes of this Case Study, just the BL's Web Archiving operations will be considered.

6.2. Organisation and activity

6.2.1. Staffing

The BL Web Archiving Team is responsible for managing, running and performing the Web Archiving activity. Selection of materials is performed by a panel of experts from across the BL. An external contractor (Magus) provides hosting for both the Web Archiving software and the gathered materials.

The BL Web Archiving team consists of:

Web Archiving Programme Manager
Technical Engineer
Curator, Web Archiving
UKWAC Project Manager
2 Web Archiving Officers
Rights Management Officer

6.2.2. Software

The PANDAS[2] software, developed by the National Library of Australia[3], is used in conjunction with the HTtrack[4] to manage and perform the Web Archiving. Details of the archived web holdings are recorded in a separate database.

[2] http://www.nla.gov.au/nla/staffpaper/2001/cathro3.html
[3] http://www.nla.gov.au/
[4] http://www.httrack.com/

6.2.3. Timing and activity

The BL established the Web Archiving team in April 2004. Following a lengthy setup phase it began archiving web sites in earnest in 2005. It currently archives around 1000 web site instances per year.

The Web Archiving activities of UKWAC are considered to be experimental and it is assumed that as experience is developed, software becomes more stable and gathering activity is scaled up, then efficiency will dramatically increase and costs will come down. It is important to bear this in mind when considering this Case Study and the wider implications of costing Web Archiving.

6.2.4. Gathering cost data

The Web Archiving team gather a range of statistics detailing their activities in the course of their everyday work, which proved invaluable for the purposes of this study. In some areas it was deemed necessary to capture additional or more detailed costings data. In particular, the Web Archiving Officers and Rights Management Officer recorded a detailed breakdown of their activities during October 2005. These activities were classified by the Web Archiving team and were mapped to the LIFE model elements.

It should be noted that gathering this data was not an easy process and should therefore be considered to be a good indication of staff effort if not a completely accurate recording. For example, the permissions process often involves sporadic effort spread across a number of weeks or even months for a single title. It was therefore difficult to account for effort spread across a number of months. It is clear from the raw data gathered that inaccuracies were present in the activity logging process.

Effort performed by staff was recorded in minutes or as an estimated percentage of their work time. These efforts were converted into costs by utilising work time and average per grade salary costs obtained from Human Resources at the BL. All staff involved in Web Archiving is currently based in London.

6.2.5. Scope of costings

Determining the scope of what is classified as a direct Life Cycle cost and what is outside of the scope was not at all clear, and this was considered an exploratory process for the LIFE Project. It was felt that only direct costs should be included, with support, management, communication costs falling outside the scope. The key requirement for LIFE was to ensure consistency across the three Case Studies. The appropriateness of the selected scope could then be debated and possibly developed in further work. Support, management, communication and setup costs were not included in this exercise.

6.2.6. Web Archiving: Titles and Instances

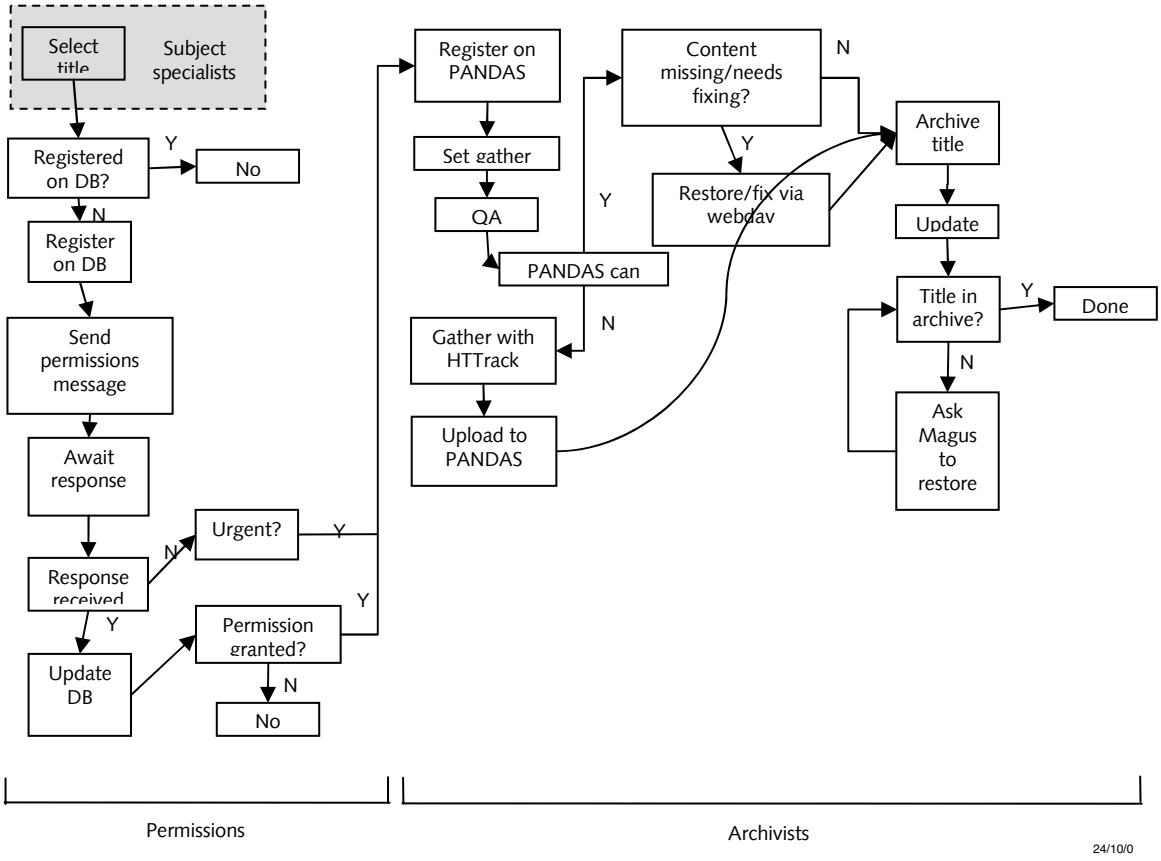
The BL's Web Archiving process can be broken down into a number of distinct steps. It should be noted that for each web site that is *selected* for archiving, a *frequency* with which that site will be gathered and archived will also be chosen. Each time that site is then gathered, a new *instance* will be archived. Costings

have been carefully recorded as title or instance related costs to enable more useful analysis of the results. Overall costings were obtained by combining per title costs with a number of instance costs per year. Statistics for the average target frequency of instance gathering (just over 4 per year) enabled overall costs to be calculated and presented in a reasonably meaningful way.

During October 2005, 141 archived instances were collected, 133 new titles were selected and 17 new titles were archived.

6.2.7. The Web Archiving process: an overview

The diagram below provides a visual representation of the Web Archiving workflow:



Beginning the process, a group of subject specialists select a list of web sites or titles to be gathered. This is then passed to the Web Archiving team. The title is registered in the database to enable it to be tracked throughout the process. Research is then performed by the Rights management officer to find appropriate contact details for the each title. A request to archive the site is then sent. Appropriate action is then taken based on the response. This could involve chasing the owners further, looking for alternative contact details, or upon receipt of appropriate permission, the process will proceed to the next phase. Details of the title are registered in PANDAS and then the gather process will begin. After obtaining an instance of the title, the Web archivists will perform a visual quality assurance check, comparing the gathered instance with the original. Necessary fixes are applied, and the instance is then committed to the

archive. Storage and backup is performed by a third party, Magus. Access is also provided by Magus.

6.3. Analysis using the LIFE Project Lifecycle Model

The following sections will analyse and cost each aspect of the Web Archiving lifecycle, categorised using the following formula:

$L_T=Aq+I_T+M_T+Ac_T+S_T+P_T$
L) is the complete lifecycle cost over T) the length of time in question

This approach is described in detail in section 4 of this report.

6.3.1. Acquisition (Aq)

Selection (Aq1)

The selection process is performed by a group of subject specialists from across the BL. Additional work was also performed by the Web archivists and the Rights Management officer.

Archivist 1 effort over one month (minutes	Archivist 2 effort over one month (minutes)	RM Officer effort over one month (minutes)	Selection board 3 grade As (minutes)	Selection board 7 grade Bs (minutes)	Total cost for October 2005 Web Archiving	Average cost per new title
140	205	1132	1161	2709	£1,813.35	£13.63

Effort expended on the selection process for one month cost £1813.35, which equated to an average per title cost of £13.63.

IPR (Aq2)

Obtaining permission to gather, archive and preserve each title is performed primarily by the Rights Management Officer with limited assistance from the Web archivists.

Archivist 1 effort over one month (minutes	Archivist 2 effort over one month (minutes)	RM Officer effort over one month (minutes)	Total cost for October 2005 Web Archiving	Average cost per new title
0	15	1224	£293.82	£2.21

Effort expended on the IPR process for one month cost £293.82, which equated to an average cost of £2.21 per title.

Obtaining (Aq4)

Obtaining each instance of a title for archiving involves effort from the Web archivists to register instances of a title in the PANDAS software and set gather operations. Problems to do with the gather process need to be rectified by the team (for example, amending database entries, fixing or repeating broken gathers, addressing issues with the PANDAS gather process). In some cases effort is expended by Magus who host the gather software.

At the current time nothing further is captured. This could be expanded to include the obtaining of lower fidelity versions of the title/instance in question in order facilitate preservation and subsequent verification of the accuracy of preservation activity. Examples might include desiccated versions of gathered web pages[5], and/or bitmap screen grabs of a sample of pages from the title/instance in a current or number of current web browsers.

Archivist 1 effort over one month (minutes	Archivist 2 effort over one month (minutes)	RM Officer effort over one month (minutes)	Magus enquiries and fixes	Total cost for October 2005 Web Archiving	Average cost per instance archived (141 instances)
205	96	0	£2373.75	£2,460.99	£17.45

Effort expended on the Obtaining process for one month cost £2460.99, which equated to an average cost of £17.45 per instance.

Licensing, Ordering and invoicing, and Checking in are not found in the Web Archiving process.

6.3.2. Ingest (I)

Quality Assurance (I1)

Quality assurance is a key area of effort for the Web archivists who spend quite a large proportion of their time manually checking and fixing broken elements of gathered web pages.

Example 1 : Antipathy.org	Example 2 : Camra.org	Example 3 : electoralcalculus.co.uk	Example 4 : infed.org	Example 5 : insaph.kcl.ac.uk/ala2004/
32	106	12	17	74

[5] Kunze, John, "Practical Approaches to Future-Proofing Institutional Web Sites" http://www.dcc.ac.uk/events/fpw-2006/fpw_2006_kunze.ppt

The table above shows time in minutes expended by one of the Web archivists in performing QA on a sample of five different titles. While effort expended on these samples for other activities remained largely constant, the QA process varied considerably. As the sample of QA data shows, this could range from relatively little effort, to considerable manual intervention.

Archivist 1 effort over one month (minutes	Archivist 2 effort over one month (minutes)	Total cost for October 2005 Web Archiving	Average cost per instance archived (141 instances)
2536	3778	£1,830.01	£12.98

Effort expended on the QA process for one month cost 1830.01, which equated to an average cost of £12.98 per instance.

Deposit (I2)

The Web archivists will commit a completed instance to the archive, check this process has been completed successfully and address any resulting problems.

Archivist 1 effort over one month (minutes	Archivist 2 effort over one month (minutes)	Total cost for October 2005 Web Archiving	Average cost per instance archived (141 instances)
608	201	£234.48	£1.66

Effort expended on the Deposit process for one month cost £234.48, which equated to an average cost of £17.45 per instance.

Holdings Update (I3)

A range of activities conducted by the Web archivists ensure that details of the holdings are maintained. This includes the current permissions state of each title.

Archivist 1 effort over one month (minutes	Archivist 2 effort over one month (minutes)	RM Officer effort over one month (minutes)	Total cost for October 2005 Web Archiving	Average cost per instance archived (141 instances)
58	1404	2109	£922.51	£6.54

Effort expended on the Holdings Update process for one month cost £922.51, which equated to an average cost of £6.54 per instance.

6.3.3. Metadata (M)

Characterisation (M1)

Currently, no characterisation activity is conducted as part of the Web Archiving process. For the purposes of the LIFE Case Study, the Web Archiving team performed a basic MIME type identification process on a single months worth of gathered web site instances (38). The table below shows the results of this process.

File Format (MIME)	Number of objects in sample
text/html	222882
image/jpeg	14837
image/gif	8553
application/pdf	2283
application/msword	1520
text/plain	1219
text/css	506
text/xml	429
application/octet-stream	412
image/png	322
application/x-javascript	155
audio/x-pn-realaudio	117
audio/mpeg	96
application/x-shockwave-flash	86
video/quicktime	53
application/vnd.ms-powerpoint	47
video/x-ms-asf	40
application/xml	31
application/rdf+xml	26
application/zip	35
audio/midi	20
application/atom+xml	17
text/rtf	17
application/vnd.ms-excel	16
image/x-icon	13
video/mpeg	13
audio/x-wav	12
video/x-ms-wmv	12
application/vnd.sun.xml.impress	8
audio/x-scpls	7
audio/wav	6
video/unknown	3
audio/basic	2
application/ogg	1
video/mp4	1
application/vnd.rn-realmedia	1

As can be seen, a very large number of HTML files were present, along with considerable numbers of image files, a number of text and document type files and then much smaller numbers of assorted multimedia objects. This sample shows that each title consists of quite a large number of digital files.

Potentially this process could be greatly expanded to include detailed file format identification and verification, as well as information describing the context of the gathered title.

Descriptive Metadata (M2)

A very small amount of descriptive metadata is recorded about each title. This includes the name of the web site, the URL and contact details. Currently no other metadata is recorded and it is envisaged that a greater amount of descriptive metadata will be captured in the future.

Archivist 1 effort over one month (minutes	Archivist 2 effort over one month (minutes)	Total cost for October 2005 Web Archiving	Average cost per new title
12	237	£72.17	£4.25

Effort expended on the Descriptive Metadata process for one month cost £72.17, which equated to an average cost of £4.25 per instance.

Preservation Metadata (M3)

No preservation metadata is recorded as part of the VDEP process, but estimations are discussed under Preservation (P), below.

6.3.4. Access (Ac)

Adding and maintenances of links, Reference linking (Ac1)

The Web Archivist is responsible for maintaining the four catalogue records which represent the web materials in the BL catalogue. She was estimated to spend 5% of her time on this activity. Providing title based entries in the BL catalogue would greatly increase the cost of this exercise although it is envisaged that the core part of this work would potentially be automated.

Web archivist, grade A (minutes)	Total cost for October 2005 Web Archiving	Average cost per new title
387	£168.54	£1.20

Effort expended on the Adding and maintenances of links, Reference linking process for one month cost £168.54, which equated to an average cost of £1.20 per instance.

Access Mechanism (Ac2)

As with many of the BL's collections, access is provided only to a very small number of the vast range of items held. Access to the archived web sites is estimated to be 5% of Magus's hosting charges (the remaining 95% covering

hosting of the Storage). It is expected that as the Web Archiving activities grow, the access costs will also increase.

Jan 05 to Dec 05 Actuals														
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total	Average Mthly
Hosting	1200	1200	1200	1200	1200	1750	1750	1750	1750	1750	1750	1750	18250	£1,521
Support	3210	2100	1650	1875	5775	2175	1650	1725	4275	1125	1500	1425	28485	£2,374

Charges made by Magus varied depending on the activities throughout the year, so average costs throughout 2005 were used.

5% of Magus hosting	Average cost per instance archived (141 instances)
£76.04	£0.54

Effort expended on the Access mechanism for one month cost £76.04, which equated to an average cost of £0.54 per instance.

There were no User Support (Ac3) costs associated with Web Archiving.

6.3.5. Storage (S)

Magus provides bit stream storage under contract to the BL. This includes facilities for backup and restore. Magus provides a breakdown of costs for Hosting and Support, as shown above under Ac3. The Hosting cost is split between storage and access and is estimated at 95% and 5% respectively (sourced from the Web Archiving Team).

95% of Magus hosting	Average cost per instance archived (141 instances)
£1,444.79	£10.25

Effort expended on the Storage mechanism for one month cost £1444.79, which equated to an average cost of £10.25 per instance.

6.3.6. Preservation (P)

No preservation activity is currently performed on the gathered web materials. For the purposes of this Case Study (and the other Case Studies undertaken by

LIFE) a generic model was developed to estimate the costs of preservation over time. See Chapter 8, The Generic LIFE Preservation ModelThe Generic LIFE Preservation Model, for more details on the model itself and how it was devised.

Applying the model to the Web Archiving Case Study involved a great deal of estimation and so the results should be considered only a very rough guide to the potential costs of preserving web materials. However, they do raise some interesting issues which can be explored in further work.

The model is based around the preservation costs for numbers of files of specific file formats. As no Characterisation work is performed by the Web Archiving team, LIFE could only utilise some estimated data calculated from a sample of simply characterised objects (a single months worth of gathered instances). In order to cost a realistic quantity of objects, the averaged sample was scaled to represent the 1000 instances that the Web Archiving team are currently aiming to gather every year.

	1 Year	5 Years	10 Years	20 Years
Total cost	£493,169	£915,219	£1,288,332	£1,617,541
Average cost per instance	£493.17	£183.04	£128.83	£80.88

As shown under Characterisation above, each archived instance consists of a large number of digital objects and consequently the estimated preservation costs are very high. Over time, the costs will fall as (in theory) more funding is put into the development of preservation tools, and file formats gradually become more stable. This is reflected in the average cost per instance which falls considerably from £493.17 in the next year of preservation to the average cost per year over the next 20 years which is a much lower £80.88.

It is clear that preserving web materials accurately but cost effectively will be a considerable challenge. The BL Web Archiving team themselves have suggested that the high volume of web materials will necessitate on demand preservation techniques where the costs are less dependant on the number of objects being preserved. This will be an exciting and crucial area of research and development over the coming years.

For the purposes of the overall Web Archiving Case Study costings, the average cost per instance across the 20 year time frame has been utilised.

Preservation element	1 year	5 year	10 year	20 year
Technology watch (P1)	5%	12%	17%	28%
Preservation tool cost (P2)	61%	53%	45%	24%
Preservation metadata (P3)	5%	5%	5%	7%
Preservation action (P4)	15%	15%	16%	21%
Quality assurance (P5)	15%	15%	16%	21%

Analysis of the breakdown of the Preservation costs provides some interesting conclusions. Investment is clearly needed in preservation tools (migration, emulation, etc) in order to lessen load on institutions establishing preservation solutions. Over time, the LIFE model suggests spending on tools will lessen if this investment is made. Costs can be lowered if the process of integrating tools within an institution's preservation and repository workflow can be simplified.

Performing technology watch, selecting appropriate preservation strategies, and recording preservation metadata (representation information) are ripe for service provision. Sharing the cost across a number of subscribing institutions will reduce costs considerably.

Reducing the costs of preservation action may be possible where on demand rendering techniques can be utilised. Quality assurance however, is harder to avoid completely. Costs can be saved by investment in automated verification and validation tools, providing effective error reporting in the preservation tools that perform preservation actions themselves, and by utilising testbeds and certification services to indicate the confidence level in preservation action tools and hence the necessary level of QA.

6.4. Overall costings

Full details of the Web Archiving Case Study can be found in the accompanying spreadsheets, but a summary of the overall costings is provided below:

Category	Percentage of overall cost (10 year average)	Average cost per instance archived	Average cost per new title	Cost per title after 1 year	Cost per title after 5 years	Cost per title after 10 years	Cost per title after 20 years
Aq	14%	£17	£16	£108	£475	£934	£1,852
I	16%	£21	£0	£111	£557	£1,114	£2,229
M	0%	£0	£4	£4	£4	£4	£4
Ac	0%	£1	£1	£4	£15	£30	£57
S	8%	£10	£0	£54	£270	£539	£1,078
P	62%	£81	£0	£426	£2,127	£4,255	£8,509
Total	100%	**£130**	**£21**	**£707**	**£3,449**	**£6,876**	**£13,731**

The per title costs for 1, 5, 10 and 20 years are based on the average cost per title, combined with the cost of gathering a number of instances of that title. On average the Web Archiving team aims to gather just over 5 instances of each title per year. In reality titles are gathered at different frequencies depending on the nature of the title in question. These figures do not include numbers for web sites which close or remain unchanged.

The percentage of overall costings reveals some interesting issues. Although the Preservation costs are estimated it is not surprising that for web materials the cost is expected to be high in relation to the other areas of activity. A likely solution to this challenge is the development of automated tools and processes which can be used to address the large volumes of materials. The Ingest and

Acquisition costs remain significant in relation to preservation and are likely to remain so. The majority of the activities conducted in these areas cannot be automated and so are likely to remain costly over the medium term at the very least.

Metadata and Access costs are insignificant at current levels but are both likely to increase as the Web Archiving process becomes more comprehensive and more useful to users. Automation of these processes will however hopefully maintain these costs at a fairly low level in relation to the other areas of activity.

6.5. Further work

Web Archiving at the British Library is in its infancy and so repeating and developing the techniques used in this Case Study will provide more useful costing data in the near future. A crucial area which has not been covered is Web Archiving domains rather than just selected sites.

7. UCL e-journals Case Study

Authors : Paul Ayris and James Watson

7.1. Introduction

7.1.1. UCL background

UCL's student base

UCL is a research-intensive university in the centre of London. In terms of student numbers, UCL has seen a steady growth during the last decade to 19,299 in the academic year 2005-06.[6]

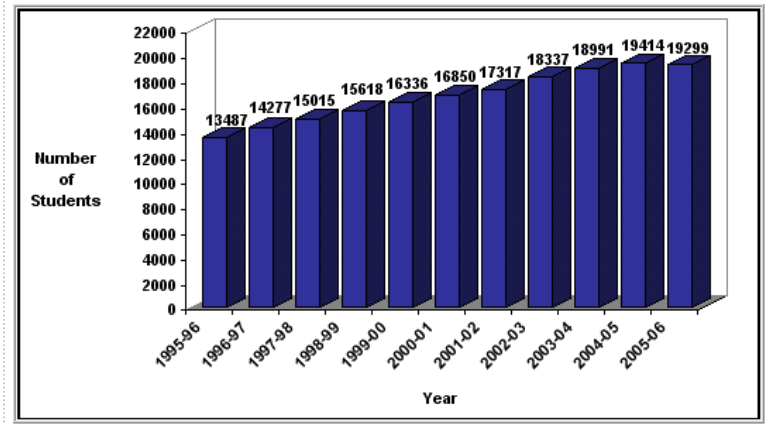


Table 1: Student Numbers 1995-2005

Year	Undergraduate			Graduate			Total		
	Full Time	Part Time	Total	Full Time	Part Time	Total	Full Time	Part Time	Total
2005-06[1]	11894	190	12084	4737	2478	7215	16631	2668	19299
	98%	21%		66%	34%		86%	14%	

Table 2: Student Numbers by category of student 2005-06

The spread of academic courses is very extensive. UCL is a multi-faculty university with a particular strength in clinical and pre-clinical studies.

Department	Undergraduate			Graduate				Grand Total
	Full Time	Part Time	Total	Full Time	Part Time	Study Leave[2]	Total	
Anatomy and Developmental Biology	141	1	142	101	17	2	120	262
Anthropology	209	0	209	110	17	21	148	357

[6] See http://www.ucl.ac.uk/registry/statistics/contents/diagram/?diagram=1&year=05. From 2003-04, the figures include Language Centre students on non-degree programmes.

Institute of Archaeology	216	1	217	241	88	3	332	549
Bartlett School of Architecture	631	8	639	351	273	3	627	1266
Biochemical Engineering	118	0	118	97	10	1	108	226
Biochemistry and Molecular Biology	187	3	190	101	4	0	105	295
Biology	214	2	216	61	7	0	68	284
Centre for Advanced Instrumentation Sys.	0	0	0	0	2	0	2	2
Centre for European Studies	0	0	0	37	14	2	53	53
COMPLEX	0	0	0	36	0	0	36	36
CALT	0	0	0	0	100	0	100	100
Chemical Engineering	149	2	151	32	7	0	39	190
Chemistry	269	1	270	106	4	1	111	381
Civil and Environmental Engineering	211	0	211	47	6	0	53	264
Computer Science	208	1	209	199	17	0	216	425
Development Planning Unit	0	0	0	100	15	7	122	122
Dutch Language and Literature	32	2	34	29	9	0	38	72
Earth Sciences	144	1	145	55	21	1	77	222
Economics	554	11	565	129	8	0	137	702
Electronic and Electrical Engineering	178	0	178	130	191	0	321	499
English Language and Literature	277	0	277	56	11	0	67	344
European Social and Political Studies	148	1	149	0	0	0	0	149
Slade School of Fine Art	151	0	151	114	3	0	117	268
French	229	0	229	10	5	1	16	245
Gatsby Computational Neuroscience Unit	0	0	0	9	0	0	9	9
Geography	338	2	340	110	19	3	132	472
Geomatic Engineering	32	0	32	57	13	0	70	102
German	154	0	154	4	2	0	6	160
Greek and Latin	153	4	157	22	14	0	36	193
Hebrew and Jewish Studies	31	8	39	14	18	0	32	71
History	320	2	322	72	22	0	94	416
History of Art	168	2	170	48	14	0	62	232
Human Communication Science	205	8	213	67	9	0	76	289
Human Sciences	112	0	112	0	0	0	0	112
Italian	196	7	203	38	13	0	51	254
Language Centre (Affiliates)	87	0	87	0	0	0	0	87
Laws	560	0	560	393	44	0	437	997
Library, Archive and Information Studies	91	1	92	78	79	0	157	249
Management Studies Centre	0	0	0	11	9	0	20	20
Mathematics	480	11	491	32	0	0	32	523
Mechanical Engineering	268	11	279	62	51	0	113	392
Medical Physics	19	1	20	56	41	0	97	117
Natural/Physical Sciences	25	0	25	0	0	0	0	25

Pharmacology	108	3	111	25	2	0	27	138
Philosophy	146	0	146	37	8	0	45	191
Phonetics and Linguistics	104	0	104	63	19	0	82	186
Physics and Astronomy	333	70	403	102	5	1	108	511
Physiology	293	4	297	37	84	0	121	418
Political Science	0	0	0	165	51	0	216	216
Psychology	357	3	360	261	130	0	391	751
Scandinavian Studies	101	2	103	13	7	1	21	124
School of Slavonic and E. European Stds	556	6	562	82	40	1	123	685
Science and Technology Studies	66	1	67	6	5	0	11	78
Space and Climate Physics	0	0	0	31	2	0	33	33
Spanish and Latin American Studies	128	0	128	7	7	0	14	142
Statistical Science	203	5	208	36	5	0	41	249
Pre-Clinical Studies	673	0	673	0	0	0	0	673
Clinical Studies	1175	1	1176	2	28	0	30	1206
Wolfson Inst. for Biomedical Research	0	0	0	13	6	0	19	19
Institute of Child Health	0	0	0	103	121	5	229	229
CHIME	23	0	23	0	78	0	78	101
Clinical Neurosciences	0	0	0	0	2	0	2	2
Ear Institute	58	1	59	19	44	0	63	122
Eastman Dental Institute	0	0	0	110	225	1	336	336
Epidemiology and Public Health	0	0	0	34	14	1	49	49
Gynaecological Oncology	0	0	0	3	0	0	3	3
Haematology	0	0	0	9	10	0	19	19
Immunology and Molecular Pathology	18	0	18	52	23	0	75	93
Infection	0	0	0	15	37	0	52	52
Medicine	0	0	0	59	60	0	119	119
Mental Health Sciences	0	0	0	4	49	0	53	53
RLW Institute of Neurological Studies	0	0	0	0	2	1	3	3
Institute of Neurology	0	0	0	74	63	0	137	137
Nuclear Medicine	0	0	0	2	0	0	2	2
Obstetrics and Gynaecology	0	0	0	15	13	0	28	28
Oncology	0	0	0	22	3	0	25	25
Institute of Ophthalmology	0	0	0	25	19	0	44	44
Institute of Orthopaedics and Musculoskeletal Science	18	0	18	7	10	0	17	35
Paediatrics and Child Health	0	0	0	5	3	0	8	8
Pathology	0	0	0	6	4	0	10	10
School of Podiatry	4	3	7	0	0	0	0	7
Primary Care and Population Sciences	22	0	22	6	33	2	41	63
Surgery	3	0	3	14	43	0	57	60
Institute of Urology	0	0	0	8	38	0	46	46

Grand Total	**11894**	**190**	**12084**	**4687**	**2470**	**58**	**7215**	**19299**

Table 3: Departmental Student Numbers by Method of Study 2005-06

UCL research

The Government's 2001 RAE (Research Assessment Exercise) awarded top marks of 5 or 5* to 58 UCL Departments.

Table 4: UCL's 5, 5 and 'best 5*' Departments in the 2001 RAE*

Of these 58 Departments, 15 were later classified by HEFCE (Higher Education Funding Council for England) as the 'best 5*' for HEFCE research funding purposes. UCL's 58 top-rated Departments included more than 1500 full-time equivalent academic staff who were entered as research active.[7]

UCL Library Services

UCL Library Services was founded in 1829, some three years after the foundation of University College London itself in 1826. The SCONUL statistics for 2003-04 and 2004-05 provide the following snapshot of the activity within UCL Library Services.

Measurement	Metric
Annual Library budget as % of total institutional annual budget	4%
Collection Size:	
▪ Catalogued Books	1,902,514
▪ Metres of archives and manuscripts	2,535
▪ Periodicals (Current subscriptions)	12,365
Ratio of spend on library staff to all other library expenditure	1 : 1.11
Ratio of spend on library staff to all other library collections	1 : 0.92

[7] See http://www.ucl.ac.uk/images/annualreport0304.pdf.

Library spend per FTE student	£581
Library spend per user	£322
Ratio of internal to external registered library users	71:29 (2004-05)
Total Staff Expenditure	£4,834,875
Total Other Expenditure	£5,378,564
Total Library Collection	4,465,325 items
Total Student FTE	17,583
Registered External Users	14,096

Table 5: Selected statistics from UCL's return to SCONUL 2003-04 and 2004-05

A number of conclusions can be drawn concerning the nature of the library service which UCL provides. First, the number of current periodical subscriptions, on both paper and e-formats, at 12,365 is high for a UK Higher Education institution and reflects both UCL's research-intensive activity and its research-led teaching. Second, the percentage of the total library budget spent on staffing is 47.34%. This is low for a CURL institution, a research-led library in the Consortium of Research Libraries. The average spend on staffing in a CURL library is not accurately known, but is commonly supposed to be at least 50% of total library budget. Third, the Library has a very high percentage of external registered users to the total number of registered library users. This underlines UCL's role as an important national collection for research purposes, but with the consequent challenges of providing electronic access to users who are not themselves members of the UCL community.

7.1.2. UCL's strategic development

UCL has a transparent roadmap for its future development, following Provost Malcolm Grant's White Paper on the Future of UCL in 2004.[8] The vision, which the Provost has propounded for UCL, is for:

- the development of UCL as 'London's Global University'
- the pursuit of international excellence in all areas of academic activity
- the need for a strategy for change management on a university-wide scale
- the importance of UCL being proactive in relation to change

UCL is developing a raft of interlinked strategies to deliver the Provost's vision for his institution.

[8] See http://www.ucl.ac.uk/images/whitepaper.pdf.

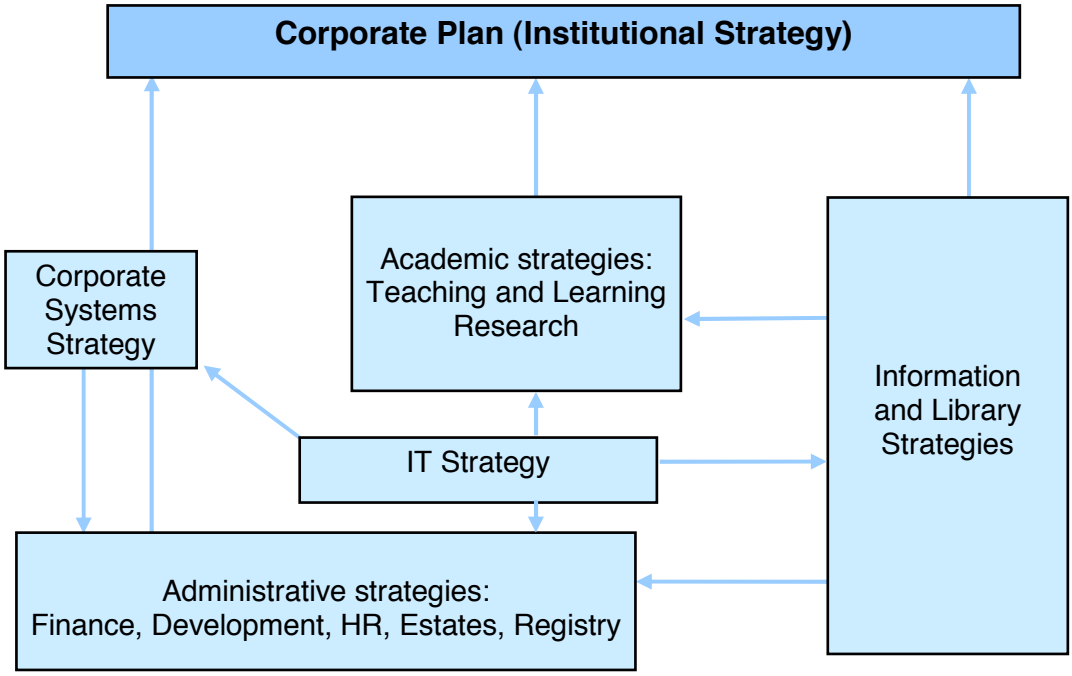


Figure 1: Inter-linked UCL strategies

The diagram above shows the joined-up nature of UCL's strategies in terms of delivering the annual Corporate Plan. Central to UCL's activity is the raft of academic strategies which it is developing for learning, teaching and research. It is important to note that, at the heart of UCL, the strategies which create the context for all other strategies are the academic strategies. All other strategies support the academic outreach of the University. There is the IT Strategy, which underpins the delivery of academic excellence in terms of hardware and software. There is the Corporate Systems Strategy, which dictates how all the corporate systems inter-relate and interact. There is then a raft of administrative strategies, which underpin the rest – estates, human resources, finance, development and fundraising. The Library and Information Strategies appear on the right of the diagram and feed into all the other top-level strategies in UCL.

7.1.3. The Library Strategy 2005-10

To support UCL's new strategic directions, UCL Library Services has itself produced a five-year Strategy.[9] The Strategy has identified ten key objectives:

1. Learning and Teaching Support (paras. 13-25)
2. Research Support (paras. 26-64)
3. Supporting the Student Experience (paras. 65-85)
4. E-Strategy Development (paras. 86-94)
5. Widening Access and Participation (paras. 95-99)
6. Fundraising Activities (paras. 100-104)
7. Partnerships (paras. 105-112)
8. Developing Library Services' Profile outside UCL (paras. 113-122)

[9] See http://www.ucl.ac.uk/Library/libstrat_may05.shtml.

9. Continuing Staff Development (paras. 123-129)
10. Planning, Resourcing and Communication (paras. 130-135)

The Library's objective for e-journals can be found in section 2 on Research Support. The Library is particularly concerned to identify a solution for the long-term archiving of e-journals, which is one of the drivers behind its sponsorship of the LIFE Project.

- In terms of Science, Technology and Medicine, provision will be made by UCL Library Services in partnership with academic Departments and will usually be in e-formats (para. 27).
- In the context of the Research Information Network, the developing role of the British Library in the legal deposit of e-materials under copyright, and commercial providers, UCL Library Services must examine its role as a paper and electronic archive and identify the correct role for itself which will support research, learning and teaching at UCL (para. 40)
- In terms of digital curation, this work will be led by the Library's Preservation Officer who, building on international good practice, will identify the role for UCL Library Services in this area (para. 41).
- As an urgent priority, research libraries and the British Library are working together to identify future responsibilities for archiving e- and paper products. In the former case, this work is being undertaken in collaboration with the JISC's Journals Working Group (para. 44).
- A Preservation Policy has been compiled by Library Services (see http://www.ucl.ac.uk/Library/preserve.shtml) which outlines the context in which preservation activities are undertaken by the Library (para. 55).
- The Preservation Policy will be reviewed, to address more explicitly issues around the digital curation of the Library's e-resources (para. 56).
- The Library will develop and publish an explicit Strategy for Preservation, showing how the aims and objectives of the policy will be implemented (para. 57).

7.1.4. The Library's E-architecture

UCL Library Services takes electronic delivery to the desktop as a given. This mode of delivery is particularly important to UCL Library Services, as the UCL family of libraries is distributed across 17 sites.[10] All but essential duplication of paper copy has been eliminated across the sites. All electronic delivery is centralised through the UCL Main and Science Libraries in the Bloomsbury campus. The Library currently has 12,365 current subscriptions to periodical titles and buys in parallel paper and electronic formats.

UCL has one of the largest, if not *the* largest, Medical School in Europe, based on three campuses – the Bloomsbury campus, the Royal Free campus and the Whittington campus. All its medical libraries are currently joint libraries serving both Higher Education and NHS users, such as nurses and PAMS, those in Professions Allied to Medicine. Largely for this reason, UCL Library Services has not yet opted for an e-only solution for the delivery of research periodical

[10] For all the library sites, see http://www.ucl.ac.uk/Library/sites.shtml.

literature. Technical and licencing barriers mean that NHS personnel, who lack honorary contracts with UCL, cannot access UCL's e-journals remotely, although they *can* have walk-in access to the majority of UCL's electronic journals at workstations in each library.

In terms of E-Strategy, this is currently developed by the Director of Library Services himself. UCL Library Services has developed a complex architecture to support its E-Strategy, which is given in Figure 2 below.

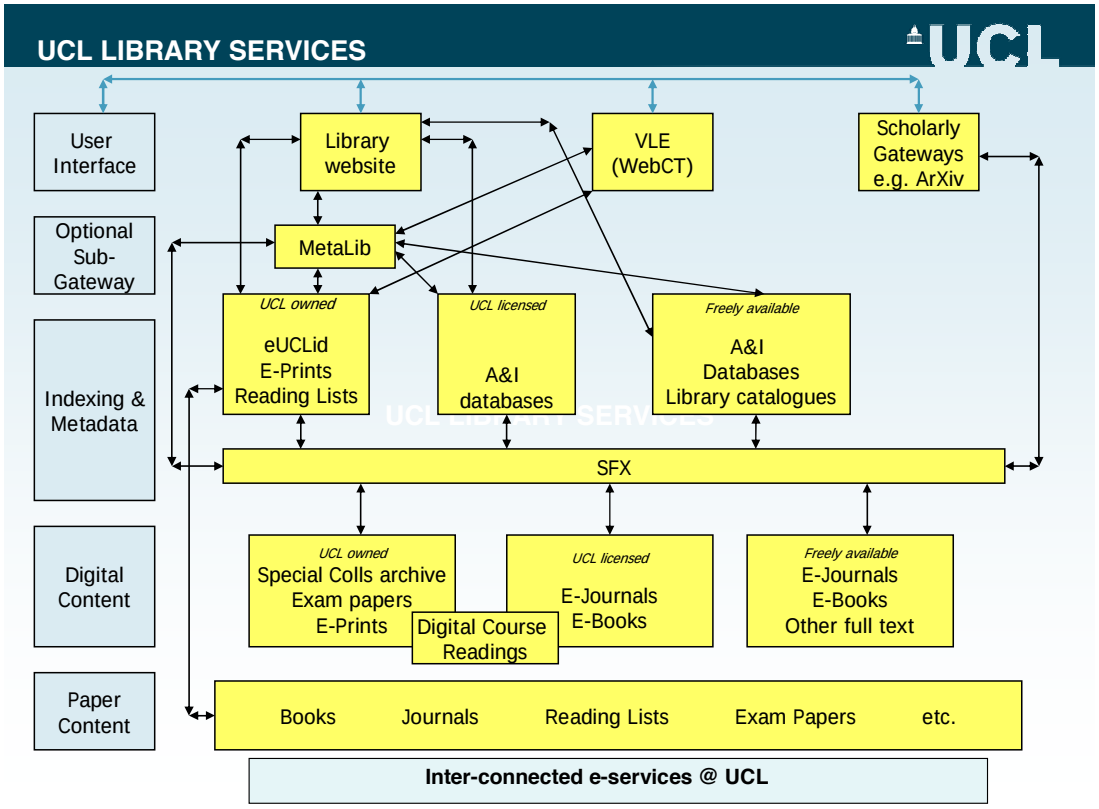


Figure 2: UCL Library Services' E-architecture

The Library's E-architecture comprises five separate levels, which are indicated in the blue boxes on the left-hand side of the diagram.

Level 1: Interface

At the top lies the user interface, the means by which users get to see the Library's content. There are a number of ways in, and the Library has deliberately made the entry points as many and as flexible as possible. The Library's website at http://www.ucl.ac.uk/Library is the most popular entry point, and the Library's e-journals are currently listed here at http://www.ucl.ac.uk/Library/ejournal/index.shtml. There are other entry points: principally WebCT, which is being used as the university's platform for its Virtual Learning Environment. Through the Library's Subject Librarians, and through the Teaching and Learning Support Section (TLSS), academics can add links to e-journal content into their WebCT courses where UCL maintains a

subscription. There are also Open Access scholarly gateways, such as the Physics archive arXiv, which are freely available on the Internet.

Level 2: MetaLib

The second level in the diagram shows MetaLib, the gateway software from the Library's supplier Ex Libris. MetaLib has recently been installed in the Library and was launched in September 2005. MetaLib allows users to search a range of databases and resources in **one** search and to be taken to the full-text of the resulting material, where this has been enabled by the SFX linking tool. MetaLib can be found at http://www.ucl.ac.uk/library/metalib.shtml. It is an extremely powerful tool and is an important addition to UCL's digital library service.

The figures below give indications of the number of MetaLib logins at representative weeks (the second week) of October, November, December and January (2005-06). Fluctuations according to the point in the academic terms should be noted.

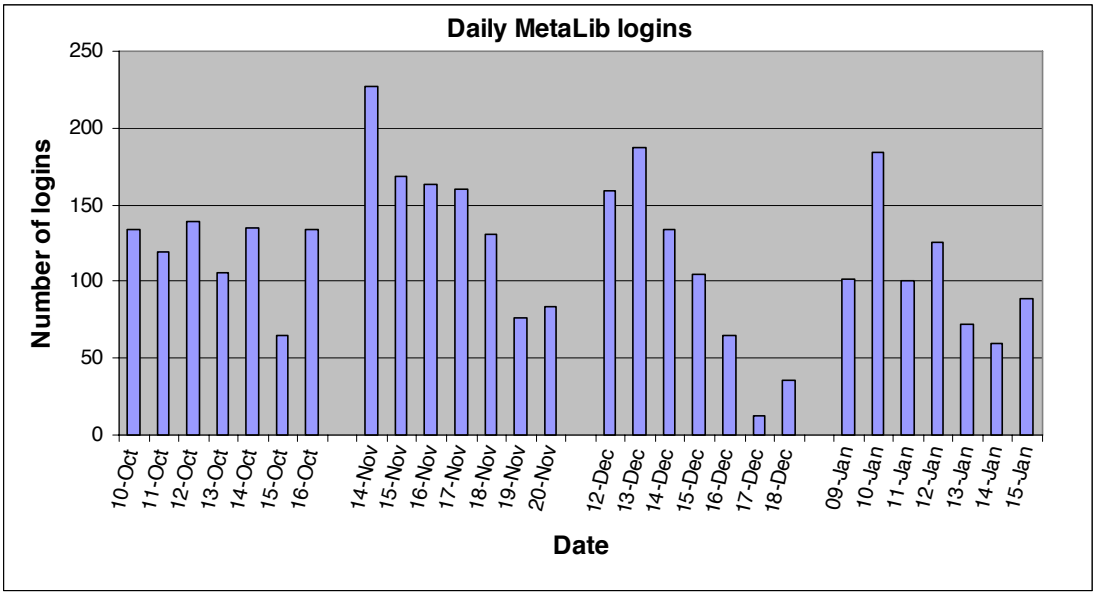


Table 6: Sample use statistics for MetaLib

Table 7 shows the number of registered users (users who have logged in at least once), listed by department for the 20 departments with the highest number of users. The total number of users registered to date is 1350. The increase of users in each department from October 2005 to January 2006 is shown. These data are limited. No adjustment is made for each department's total numbers of staff/student members, and it is not possible to obtain numbers of logins or the amount of use per department, as the statistical data for logins/searches is anonymized. Neither is it possible to distinguish staff from student users. In terms of the number of users trying MetaLib at least once, each Faculty is represented in the top 20.

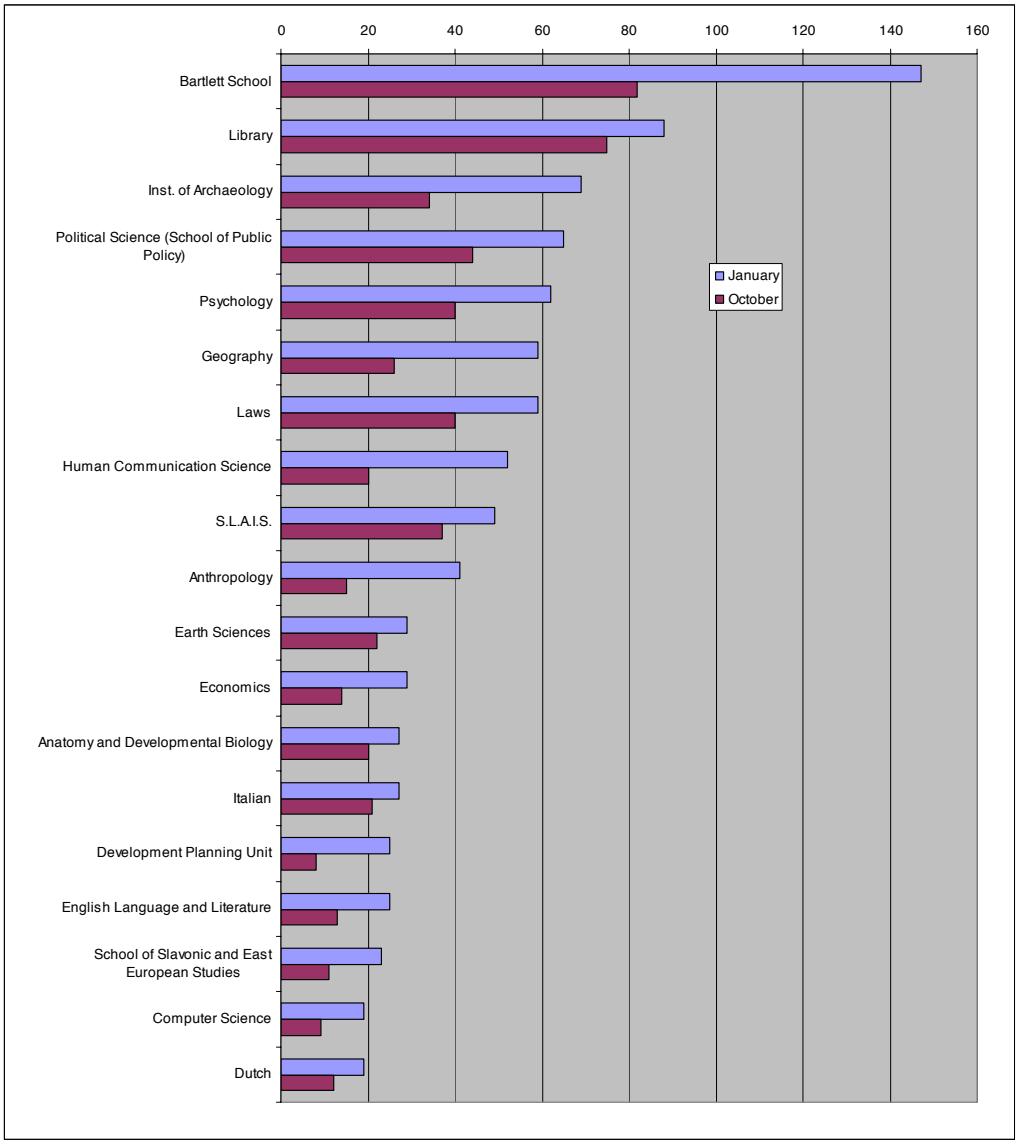


Table 7: Use of MetaLib by academic Department

Level 3: Indexing and Metadata:

The third level in the Library's E-architecture, represented in Figure 2 above, is Indexing and Metadata. This can take a variety of forms: the Library catalogue eUCLid, which is Aleph from Ex Libris, Reading Lists or Abstracting and Indexing databases. SFX is again an extremely important component in this level. Again supplied by Ex Libris, SFX links the user from a reference which he/she has found directly to the most appropriate e-copy to which he/she has rights to access.

Level 4: E-content and Level 5: Paper content

The fourth and fifth layers in the E-architecture are content, both digital and in paper format.

Item	Metric
Catalogued Books	1,902,514
Metres of archives and manuscripts	2,535

Periodicals (current subscriptions)	12,365

Table 8: Metrics indicating the size of UCL collections

The total collection in the Library amounts to 4,465,325 items. It is here that the Library's e-journals should be counted. In supporting research in Science, Technology and Medicine in UCL, the Library's suite of e-journals forms a major new development in the portfolio of services which UCL Library Services has developed over the last 10 years, in common with all other research-led universities in the UK.

7.2. Analysis of the UCL e-journal lifecycle

The following sections will analyse and cost each aspect of the e-Journals lifecycle, categorised using the following formula:

$$L_T = Aq + I_T + M_T + Ac_T + S_T + P_T$$

L) is the complete lifecycle cost over T) the length of time in question

This approach is described in detail in section 4 of this Report.

7.2.1. Acquisition (Aq)

UCL acquires single titles, NESLi2 packages and non-NESLi2 packages

- Subject Librarians make recommendations for single titles/are consulted on potential acquisition of new packages, and usage figures are consulted if available
- Following approval by Subject Librarian, e-journals staff:
  - Check if title is available electronically if necessary
  - Find out how much electronic access will cost (e.g. requesting quote from publisher for custom package, costing NESLi2 offer)
  - Check if publisher's definition of a "site" matches that of UCL and if proxy server is permissible, and negotiate with the publisher over price/terms if necessary
- If satisfied, title/package is placed on the wish list
- When funds are available, order is placed either with agent, publisher or Content Complete as appropriate
- Prior to payment being made, full licence terms and conditions are checked and licence signed. Licence filed for future reference
- Order and invoice are entered on Aleph
- E-journals staff monitor for access becoming available and activate as necessary – this usually involves online activation with a customer number and creation of administrator account where contact and IP information are added
- New titles are entered in Access database and SFX and publicised to users

Renewals

- Confirmed once yearly

- If renewal of an existing package (e.g. NESLi2), current year's contents listing is compared against Access database and updated as necessary. This may mean that titles have been removed from package, in which case go to Acquisition stage
- Licence terms and pricing may change so re-negotiation of price/terms may be required
- Licence is signed and checked and invoices are added to Aleph
- Increased time spent troubleshooting access problems at the time of year of the UCL Case Study

7.2.2. Ingest (I)

There is no Ingest stage as none of the content is stored locally. All content is accessed from stored content on publishers' and aggregators' servers.

7.2.3. Metadata (M)

- If title has an existing catalogue record, holdings information for electronic version is added
- If no catalogue record, basic catalogue record of at least title and ISSN is added, as well as holdings information
- Catalogue records are updated to reflect title changes etc. and holdings data is updated as necessary

No preservation/structural metadata is added as UCL does not store/preserve e-journal content.

7.2.4. Access (Ac)

- Licence negotiations/conditions will have established who is to have access. The necessary information is added to the Access database regarding use of proxy server and walk-in access
- E-journals staff activate/re-activate access where necessary and maintain records of account usernames and passwords
- E-journals staff ensure that publishers/agents have up-to-date list of IP addresses
- Maintenance of password list for titles which do not use IP recognition
- Title is added to the e-journals Access database with the following information: ISSN, EISSN, publisher, aggregator, date added to database, notes (e.g. which package it belongs to), URL, holdings and subject classification and whether UCL holds a print subscription. This information is maintained and updated as necessary, with date of update recorded
- Titles are activated in the SFX database and thresholds are checked to ensure they match UCL holdings. Maintenance of SFX to ensure that UCL holdings are correctly displayed and linked to
- Addition and maintenance of information on web pages relating to access restrictions, terms and conditions of use, Frequently Asked Questions
- Production of news items and mailing lists to inform users of ongoing problems, changes to access procedures and new titles added

- Answering user enquiries relating to access. This can be where access has been cut off, IP addresses are incorrect, users are experiencing problems with Athens/proxy log-in
- Answering staff enquiries relating to storage and access especially by walk-in users, permission to print for ILL

7.2.5. Storage (S)

UCL does not store e-journals locally.

7.2.6. Preservation (P)

UCL does not currently preserve any e-journal content.

7.2.7. Visual overview

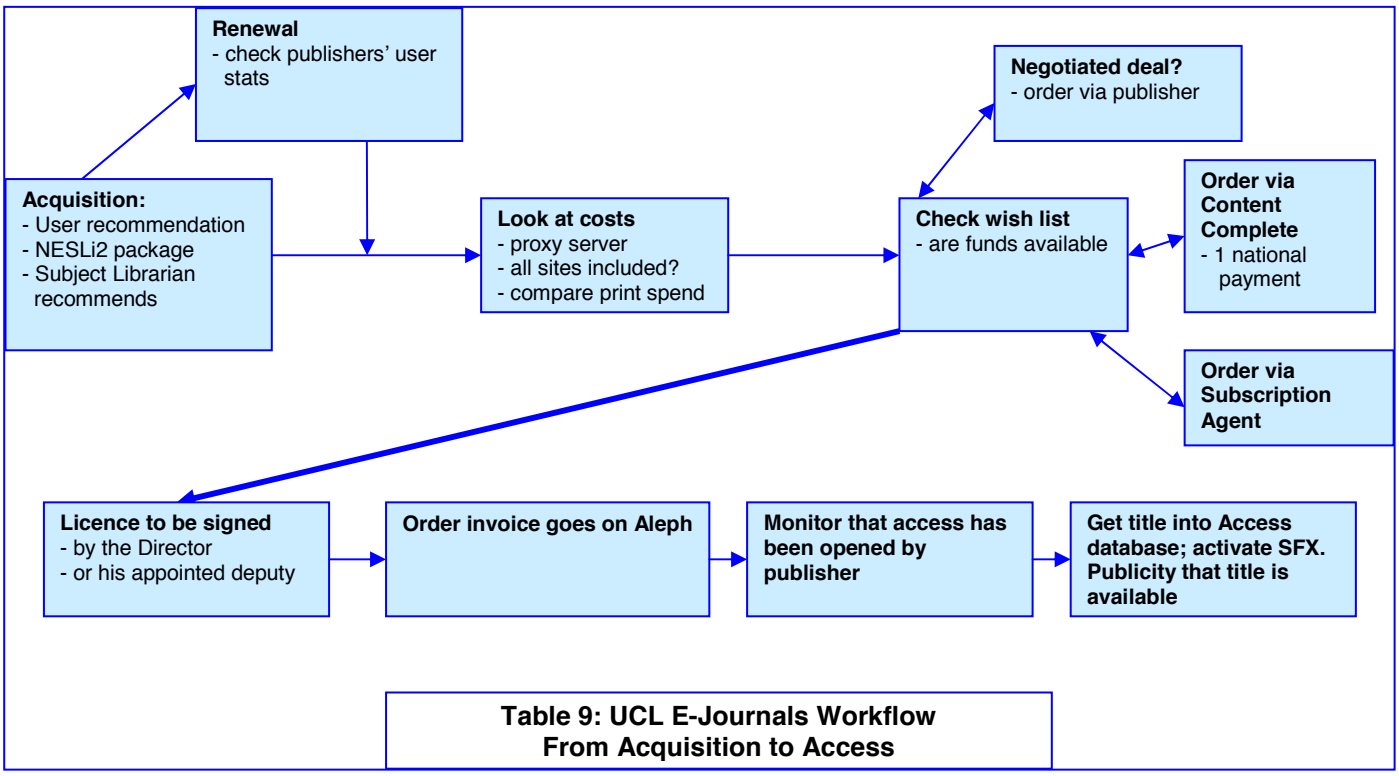


Table 9: UCL E-Journals Workflow From Acquisition to Access

7.3. Case Study examples

7.3.1. Introduction

UCL Library Services undertook two Case Studies in e-journals management as part of the LIFE Project.

This Report provides an overview and analysis of lifecycle collection management for the Public Library of Science corpus and for material from Blackwells. The lifecycle as analysed here aims to provide insight into the potential full life of the materials, within the context of the workflow of the VDEP material, the existing management of UCL e-journals and the exercises with Oded Scharfstein of Ex Libris with DigiTool v3.

This Report aims to analyse all aspects of the digital collection lifecycle of the UCL electronic journals collection. It also aims to highlight where additional processes could be implemented to aid preservation. There have been certain specific difficulties with the digital preservation lifecycle approach for the UCL e-journals. These mostly can be attributed to the fact that, as the journals operation in UCL is service-driven practically, no content is stored locally at UCL, with all content accessed over the Internet. Examples of difficulties with analysing the UCL collections include:

- It is impossible to obtain file sizes from the collection
- It is impossible to obtain breakdowns of file formats from the collection
- There is no Ingest stage in the lifecycle
- There is no Storage stage in the lifecycle
- There is no Preservation stage in the lifecycle

The titles are not generally managed in a title/issue way (i.e. when a title has been purchased there is no further check-in of single issues); this means that it is difficult to get an idea of the local storage of journals from existing procedures.

7.3.2. Case Study : PLoS (Public Library of Science)

The corpus of PLoS (Public Library of Science), as at summer 2005, comprised three journal titles consisting of some 30 issues, with more than 10,000 files of content. A further breakdown of this content appears below. Oded Scharfstein loaded exemplar files into DigiTool v.3.

The LIFE Project obtained the PLoS content on multiple DVD roms. The crucial issue for ingest is the size and shape of the files as they come in. The more files in a given object, the more complicated the ingest procedure becomes. Furthermore, the greater the human involvement in the ingest procedure (whether this is necessitated because of greater complexity, or whether it is because of standard procedures in place), the more the procedure will cost.

The average issue in the PLoS corpus contains 463 computer files. This usually consists of PDF copies of all "content" articles, with XML versions of the same articles. Each PDF copy has all pictures etc. embedded in it, each XML copy has the pictures etc. linked, with versions of the pictures as large and small TIFF files for preservation and access. Also, with the content is a large amount of supplementary information. The information in these supplementary folders is referenced in the articles but not embedded (they are still linked to from the

HTML/XML). These files vary considerably in type, from simple formats such as plain text, to highly complex and highly proprietary scientific data formats.

The complex nature of these files means that the structure of the AIP's that need to be ingested into the archive will be complex. The construction of the structural metadata needed for the information will be extensive, and it will be an involved procedure.

Some of the file formats present in the PLoS corpus are not mainstream file formats. For example: in Volume 1 Issue 2 there is extensive material which is specifically related to describing genes, and more accurately, describing the genes of nemotodes (these files have the extensions .gff, .briggsae and. elegans, the latter two names being types of nematode). To confidently add technical metadata to these files and to confidently archive these files for preservation would require good documentation concerning these file formats. As this documentation does not come with the files, this would need to come as the result of research.

How the material is supplied is also an issue here, with material that is supplied on a hand-held item costing more than material that is supplied purely electronically.

The application of catalogue records for electronic journals is similar to that for print journals. However, at the time of undertaking the LIFE Project, UCL Library Services had stopped the creation of full catalogue records for e-journals in the eUCLid catalogue because of the large number of e-journals being acquired in NESLI e-journal packages. Only basic catalogue records were created, mainly as a way to cope with invoicing and payment requirements. No technical, structural or preservation metadata is added to the e-journals at UCL because UCL does not currently preserve any e-journal content.

Access is delivered via the web and the SFX open URL resolver. Access takes up a good deal of the time of the dedicated electronic journals staff at UCL, as maintenance of links is a very large job. When accessed remotely, electronic journals are not processed on a title and issue level, there is no check-in procedure and issues are not chased systematically to check they are there. The consequence of this is that it is extremely difficult to extract a title- and issue-level costing from this exercise.

The UCL electronic journals are currently only stored remotely on publishers' and aggregators' servers. Consequently, these costs cannot be accurately ascertained from UCL.

Nor does UCL currently attempt to preserve any e-journal content locally – all the workflow in UCL, as in most UK universities, is geared towards making access available.

In the PLoS corpus the issues break down as follows:

Total number of Issues:	21
Total number of files:	9739
Average number of files per Issue:	463
Total number of "content" articles:	751
Average number of "content" articles per Issue:	35.76
Average number of files per "content"	12.96

article:	

Table 10: Breakdown of PLoS content

The journal is monthly. The average number of files in an issue is: 463. The average number of content files in an issue is about 35. The average size of an issue is: 639mb. In a particular instance of the Case Study, the files in PLoS Biology, break down as follows:

File type	Description	% of total	Number
TIFF		43.66%	3797
HTML*		17.20%*	1496*
PDF		10.26%	892
XML		9.44%	821
Excel		4.86%	423
GIF		4.67%	406
PDF		2.28%	198
Word		1.58%	137
Gpr		1.13%	98
Avi	Windows media player file	0.91%	79
EPS		0.86%	75
JPEG		0.79%	69
Txt		0.61%	53
GIF		0.56%	49
Mov		0.33%	29
PostScript		0.30%	26
Ppt		0.26%	23
Db		0.07%	6
FA	Representation of DNA information	0.05%	4
GFF	Representation of DNA information	0.03%	3
Ai		0.03%	3
Dcr		0.02%	2
Tds		0.02%	2
GZ	Compressed file created using GZIP	0.01%	1
BRIGGSAE	Scientific data on the dna of a briggsae nematode	0.01%	1
ELEGANS	Scientific data on the dna of an elegans nematode	0.01%	1
Tex		0.01%	1
DS_store		0.01%	1
Tar		0.01%	1
Total		100.00%	8697

Table 11: Breakdown of PLoS content by file type

* This figure is anomalous as 1450 files are in one issue, therefore the figures have been reproduced below with 1450 removed

As a desk study, a number of issues can be identified concerning long-term digital preservation. First, one can assess obsolescence of file formats. On examination of the file formats in the collection, it is apparent that none of the file formats in existence in the collection has become obsolete (i.e. that they are no longer accessible because they are no longer supported). However, some are

certainly aging, if still widely supported by software, a good example of this are .gif files. GIF files have been the most-widely used of web graphic formats, but they are widely deprecated because of their proprietary nature. They are also an old file format (1989) and use of the PNG (Portable Network Graphic, see: http://www.w3.org/Graphics/PNG/) has been advocated instead; they have also been used in digital preservation Case Studies as an "obsolete" file format before (see Rosenthal et al. 2005: http://www.dlib.org/dlib/january05/rosenthal/01rosenthal.html). GIF files account for about 5.26% of the collection with a total of 455 files. The VDEP archive has been in existence since 2000 and is therefore older than this material. As the VDEP collection is the oldest that we have to look at, then it seems sensible to take the figure of 5.26% every 5 years as the metric for identifying the challenges for digital preservation.

7.3.3. Blackwells data

Oded Scharfstein also considered a sample of Blackwells data in the context of DigiTool v.3. Many of the issues which have been highlighted in terms of the PLoS data are also relevant to Blackwells. The Blackwell NESLi2 package has 758 titles in 2006 and UCL estimates that this represents about 5,000 issues.

Blackwell Publishing - e-journals

Journal title	Issue or volume number	Total number of items (i.e. files)	Total size (mb)	pdf files	gif files	xml files
Allergy	57	1694	67.8	390	888	416
Allergy	60	1808	72	324	1159	324
British Journal of Haematology	112	1275	62.4	174	913	188
British Journal of Haematology	119	1361	70	187	960	214
Clinical and Experimental Allergy	31	1848	126.4	274	1300	274

	Total number of items	Total size (MB)	PDF files	GIF files	XML files
Averages	**1597.2**	**79.72**	**269.8**	**1044**	**283.2**

Table 12: Breakdown of Blackwells content

There were only three file types in the Blackwells material - PDF, GIF and XML files. The commercial material would thus be easier to handle in an e-archive

and seems to reflect a conscious decision from a commercial supplier to limit the multiplicity of file formats with which it has to deal.

In discussion with Ex Libris over the potential of DigiTool v.3 to ingest multiple file formats, the following points were noted. Regarding the LIFE Project, DigiTool does not have 'out of the box loaders' that know how to address e-journal formats, mainly because there are so many formats and there is no standard. Hence, in order to load e-journals into DigiTool, the DigiTool Product Manager manually converted them into an already-supported format, e.g. METS, for loading into the system. This is what happened with the PLoS articles. Ex Libris would be interested, once there is a customer who uses DigiTool to archive e-journals, to develop the appropriate loaders.

In the course of the Project, there was not sufficient time (due to staff changes) to convert the Blackwells data into a METS format for loading into DigiTool. The study with Blackwells data was purely a desk study.

7.3.4. Logging of tasks in UCL Library Services

UCL Library Services undertook a diary exercise for its two e-journals staff during the period 1st November 2005-14th December 2005, with particular reference to the lifecycle formula outlined above.

The lifecycle formula with which the Project is working does not, in many respects, fit the workflow for e-journal materials in UCL Library Services. UCL is geared towards **giving access** to e-journal literature, and to answering enquiries about the resulting access. The emphasis is **not** on ingest, storage nor preservation, as there is none in the strict sense of these terms.

A number of caveats need to be borne in mind about the UCL data given below.

- November was not a typical month in terms of the amount of time spent on web development. UCL introduced a proxy server and so staff spent a lot of time making changes to the way the UCL web pages look, a database is structured etc. This is something that obviously would not occur month on month
- November was also non-typical because it is the time when the e-journal administrators do most work on the UCL wish list. This means asking for title suggestions from Subject Librarians, obtaining prices, seeing if titles are available electronically etc. Time spent on this (and on actually ordering the titles) is included in the "licence checking / negotiation with publishers" figures as this is where it seems to sit best
- In light of the time UCL spent on these two major areas, the amount of time UCL spent on linking work is consequently lower than it would normally be
- The total of the figures given below does not represent the sum total of the staff's hours. They tackled other things which are not accounted for in the lifecycle tasks they were asked to record, e.g. training and current awareness, attending meetings, working on the enquiry desk, maintaining budget information and working on other projects
- Renewing titles is one of the tasks UCL was asked to record data against, but this largely falls to other members of staff outside the immediate e-

journals operation, and so does not feature in the data below
comment: Table should all be on one page]

Acquisition (Staff member A)	**Time taken**
Entering orders and invoices onto Aleph	12.50 hours
Renewing subscriptions	1.00 hours
Licence checking/negotiation with publishers	67.00 hours
Metadata (Staff Member A)	
Cataloguing	7.60 hours
Access	
Enquiry Work Staff Member A Staff Member B	 40.83 hours 16.75 hours
Linking work (SFX, Access etc.) Staff Member A Staff Member B	 11.50 hours 65.05 hours
Web development Staff Member A Staff Member B	 15.00 hours 2.20 hours
TOTAL	**239.43 hours**

Table 13: Staff activity in UCL Case Study

7.3.5. Analysis

A number of analyses can be performed on this data. It is possible for UCL to calculate, using activity-based costing, the total cost of making e-journals available to users during this period. Using the new HERA Pay Framework payscales[11] in UCL Library Services, the total cost of the Acquisition, Metadata and Access stages during the period under review was:

Staff Member	**Cost**
Staff Member A	
155.43 hours @ £17.13 per hr	£2,662.52
Staff Member B	
84 hours @ £11.68 per hr	£981.12
Total	**£3,643.64**

Table 14: Staff costs in UCL Case Study

[11] Assumes a basic working week of 36.5 hours; figures include both employer's on-costs and London Weighting.

Were these activities typical of the whole year, which they are not, it would be possible to calculate the cost of each of the three activities over the course of the year.[12] These would be as follows:

Activity	Cost
Acquisition	£10,801.89
Metadata	£1,019.81
Access	
- Staff Member A	£9,034.68
- Staff Member B	£7,685.44
Total	**£28,541.82**

Table 15: Notional Staff costs over 1 year

As well as institutional costs, the Project team tried to drill down into the data from the Case Study to look at the costs of providing access and preservation for individual journal titles using the LIFE formula for identifying these costs. The formula which the LIFE Project has identified for acquisition, storage, access and preservation of e-journals is $\mathbf{L_T=Aq+I_T+M_T+Ac_T+S_T+P_T}$.

The Project team found this to be a difficult piece of work to undertake in a Higher Education environment. There were two reasons for this:

- It is unusual for a University Library to undertake lifecycle costings and analyses of the materials that it purchases and there is no tradition of lifecycle management at the level of detail demanded by the LIFE formula
- It was not possible in the UCL Case Study to determine activity costings at a title level for the Access portion of the formula.

As will be clear from the activities listed under the Acquisition and Access stages, this is the most time-consuming portion of the work undertaken in UCL Library Services. With the number of staff available, it was not possible to break down the activities in this heading to journal package or title level and **still** deliver a robust service to UCL's users. The difficulties in recording this information need to be addressed in future iterations of the LIFE study and methodology.

The LIFE formula $\mathbf{L_T=Aq+I_T+M_T+Ac_T+S_T+P_T}$ can be broken down as follows at a title level over 1 year:

Element	Cost (Range 1)	Cost (Range 2)
Aq Purchase of titles Staff activity	 £199.72 £1.25	 £539.35 £6.23
I	£0	£0
M	£3.97	£3.97
Ac		

[12] Assumes a working year of 47 weeks.

Staff Member A Staff Member B	£1.04 £0.89	£5.21 £4.43
S	£0	£0
P	£0	£0
TOTAL	**£206.87**	**£559.19**

Table 16: Values for the LIFE formula per title over 1 year

In Aq, for purchase costs, the figures were derived from the e-only cost for two 'Big Deal' Journal packages from a UK and continental European publisher for e-only delivery,[13] divided by the number of titles in the package. The figure includes VAT. The resulting answer gives the costs of both journal packages. In Aq, the staff activity costs are an estimate. They are based on the predicted activity costs over one year divided by the number of journal titles that might experience access activity in the course of a year. The result is a range if **all** 8668 UCL e-journals were included to one fifth (or 1733 e-journals). The M cost is also an estimate, derived from the cataloguing of monographs. The Project team estimated that 25 e-journals could be catalogued in one day. Using a current costing from Bibliographic Services,[14] this works out at a unit cost per title for metadata creation of £3.97. The Access stage costs are again an estimate. They are based on the predicted costs over one year divided by the number of journals that might experience Access activity in the course of a year. The result is a range if **all** 8668 UCL e-journals were included to one fifth (or 1733 e-journals).

Based on this treatment, the answer to the LIFE formula, which the LIFE Project has identified for the acquisition, ingest, metadata creation, access, storage and preservation of UCL's e-journals over 1 year is a range of **£206.87 - £559.19** per title.

The LIFE formula is flexible and the T value allows the lifecycle costs to be predicted over any given timespan. The conclusion of the UCL Project Team was that the projection of the UCL data over a 5 and 10 year range could not be performed with any certainty due to the lack of robustness of the data from the Case Study. However, as a piece of desk research, predicted costs per title over a 5 and 10 year period were identified. These figures were worked out by taking the costs for the value T=1 year and working out *average* costs for staff and materials costs in the LIFE formula. Salary costs for years 1-10 were now predicted on the basis of the staff member being on top of grade, to show maximum costs.

Element	Cost
Aq Purchase Staff activity	 £369.54 £4.55
I	£0

[13] Based on figures for 2005 and 2006.

[14] This cost is based on the hourly rate of a member of staff on the bottom of the new HERA Pay Framework Grade 6 with on-costs, including London Weighting. The working week is assumed to be 36.5 hours.

M	£3.97
Ac Staff Member A Staff Member B	 £3.81 £2.97
S	£0
P	£0

Table 17: Average values for elements of the LIFE formula where T=1 year

For the purposes of calculation over a 10-year timeframe, an inflation factor of 7% per annum for materials costs was assumed, plus a 3.5% cost of living increase each year in staff costs. The LIFE formula was then modelled where T=5 years and T=10 years. The formula needed to be adapted to give the Aq element a value for T, because in Higher Education purchase costs for e-journal subscriptions, plus activities associated with acquisition, are an annual cost.

Element	**Yr1**	**Yr2**	**Yr3**	**Yr4**	**Yr5**	**TOTAL**
Aq Purchase Staff	 £369.54 £4.55	 £395.41 £4.71	 £423.09 £4.87	 £452.70 £5.04	 £484.39 £5.22	 £2,125.13 £24.39
I	£0	£0	£0	£0	£0	£0
M	£3.97	£0	£0	£0	£0	£3.97
Ac Staff A Staff B	 £3.81 £2.97	 £3.94 £3.07	 £4.08 £3.18	 £4.22 £3.29	 £4.37 £3.41	 £20.42 £15.92
S	£0	£0	£0	£0	£0	£0
P	£0	£0	£0	£0	£0	£0
TOTAL for Years 1-5						£2,189.83

Table 18: Predicted costs per title for years 1-5 for UCL e-journals

Element	Yr6	Yr7	Yr8	Yr9	Yr10	TOTAL
Aq Purchase Staff	 £518.30 £5.40	 £554.56 £5.59	 £593.40 £5.79	 £634.94 £5.99	 £679.38 £6.20	 £2,980.58 £28.97
I	£0	£0	£0	£0	£0	£0
M	£0	£0	£0	£0	£0	£0
Ac Staff A Staff B	 £4.53 £3.53	 £4.68 £3.65	 £4.85 £3.78	 £5.02 £3.91	 £5.19 £4.05	 £24.27 £18.92
S	£0	£0	£0	£0	£0	£0
P	£0	£0	£0	£0	£0	£0
TOTAL for Years 6-10						£3,052.74

Table 19: Predicted costs per title for years 6-10 for UCL e-journals

Values for T in LIFE formula	**Normalised costs**

T=5 years	£2,189.83
T=10 years	£5,242.57

Table 20: Predicted costs per title over 5 and 10 years

Based on the Case Study for costs where T=1 year and from desk research where T=5 or T=10 years, it is possible to posit as a hypothetical model the costs per title over this period and these are given in Tables 16 and 20. However, the LIFE formula needs to be populated with more robust data, when rigorous data collection techniques in UCL Library Services are in place, to capture the raw data needed to predict costs over a 1-10 year timespan with greater accuracy.

7.3.6. Discussion

The UCL Case Study reveals a number of features which need to be addressed in a second, fuller data capture and analysis exercise. UCL, as a research-led institution, has as its objective the acquisition of and access to e-journal content for its staff and students. At the time of the Case Study, 8668 e-journal titles were logged in the UCL Access database which generates the web listings of titles.[15] In terms of the lifecycle of the e-journal content acquired by UCL, the most significant cost is the purchase of the content itself. Unlike copyright deposit libraries, UCL has to pay for the purchase of every piece of content which it acquires. The modelling of the Aq element in an HE library also required the addition of a T element to the LIFE formula, as Aq costs are an annual cost for universities – both for subscription and staff costs.

As a service-led organisation, UCL undertakes no ingest, storage or preservation functions. Access to the e-journal content is via the remote publisher's server. Consequently, no costs for these activities appear in the Tables above. In terms of staffing activity performed on the e-journal content, the most expensive activity in terms of costs are those actions concerned with making the materials accessible to users. Indeed, all management activities in UCL Library Services are subordinate to this objective.

The logging of data in the Case Study has implications for the range of costs which can be predicted per title. During the Case Study, no log was made of the number of issues or titles which were dealt with by staff during the six-week study period. Thus, in the Tables above which are built on the 6-week data collection exercise, the staff activity costs are notional and based on a range of possible values for a notional number of titles dealt with during the trial. The weakness of the data in these parts of the formula underlines that activity-based costing is not embedded in all university libraries.

The purchase costs of two commercial e-journal packages studied in the trial are firm, but the range indicates that the values in the Tables will vary according to the package which is the subject of study. There is no such thing as a uniform purchase price for e-journal content.

The Case Studies revealed that the concept of the individual file has no function in the way UCL currently manages its e-journal content. Even individual e-issues are not logged or checked in. All activities are subordinate to purchasing the content and enabling e-access to the material at title level.

[15] See http://www.ucl.ac.uk/Library/ejournal/index.shtml.

A longer Case Study, with more refined data capture techniques, would enable the 1-year, 5-year and 10-year lifecycle costs of UCL e-journals to be ascertained with greater certainty.

7.4. Conclusions

In the original LIFE Project plan, the outcomes of the Project were intended to address the following *Key Questions* for Higher and Further Education (HE/FE):

- What are the long-term costs of preserving digital material?
- Who is going to do it?
- What are the long-term costs for a library in HE/FE to partner with another institution to carry out long-term archiving?
- What are the comparative long-term costs of a paper and digital copy of the same publication?
- At what point will there be sufficient confidence in the stability and maturity of digital preservation to switch from paper to digital for publications available in parallel formats?
- What are the relative risks of digital versus paper archiving?

The outcomes are seen as important for HE and FE for the following reasons:

- The lack of a dependable digital archive is a major inhibitor in the ability of institutions in HE/FE to move to the e-only delivery of materials
- Institutional Teaching, Learning and Research strategies are underpinned by the content which libraries provide and help to determine the nature of courses and inter-departmental collaborations
- There is a need to model the possible costs of long-term digital archiving on institutional budgets
- Institutions need to know at what stage they can stop buying parallel paper and e-formats for access and preservation, trusting to e-delivery and digital archiving alone

What information does the UCL Case Study offer in answer to these questions? As a service-led organisation, UCL performs no digital preservation activity and was unable to comment on this aspect of the LIFE formula. The lifecycle costings over 1-10 years, which UCL has identified, form the baseline against which future preservation costs can be added and measured.

In terms of identifying the responsibility for digital archiving in the future, the Case Study certainly underlined that no digital preservation was being undertaken by UCL. The Case Study, and associated work which UCL has been undertaking with Ex Libris, does suggest that UCL would be capable of undertaking such activity. The workflow in Table 9 lends itself to a lifecycle approach such as the one embedded in the LIFE formula, although with considerable additional activities needing to be included. Outside the e-journals study, UCL has led a separate evaluation of DigiTool as a platform for managing e-content. UCL led the UK programme and undertook an evaluation of DigiTool as a platform for managing e-theses. The version of DigiTool tested in that

Project was v. 2.4. whereas the present e-journals Case Study looked at Digitool v.3.

The conclusion of the earlier UCL e-theses study was that 'In a competitive market place, it is highly likely that large research libraries would want to use DigiTool for both digital asset management and digital archiving.' DigiTool v.3 conforms to the OAIS model (Open Archival Information System), and this is important with regard to an institution's ability to use this as a tool for long-term digital curation. DigiTool is not in itself a digital archive, but a set of tools and protocols which can be used to support digital archiving.

Internationally, there are a number of initiatives which are looking at the creation of long-term digital storage. In the Netherlands, the e-Depot at the Hague is carrying out long-term digital preservation using commercial e-journal content from publishers.[16] LOCKSS is another approach to digital archiving, which aims to store several copies of the same content on different servers,[17] to minimize the risk of loss. Ithaka is also looking at marketing a digital repository system.[18]

Who should be responsible for digital archiving in the UK?

- The conclusion of the UCL e-journals Case Study is that, from a technical point of view, a platform such as DigiTool could be used to manage a digital archive in an HE institution
- From a financial aspect, UCL would need to undertake more rigorous data sampling and analysis using the LIFE formulae to ascertain the costs at title and file level
- At a workflow level, UCL would need to undertake a considerable re-modelling of current practices to achieve the goal of establishing a local digital archive

The conclusions presented elsewhere in this Report show that the British Library is further advanced than HE in terms of lifecycle costings and the construction of a digital repository for the long-term storage of digital content. The present Case Study suggests, however, that digital archiving at a local level is an issue which universities should consider seriously.

In terms of identifying the comparative costs of paper and electronic archiving, the UCL Case Study did not contribute an answer to this question, as the data sampling was performed entirely on e-journal content. Nor does the study offer any information on the relative risks of paper versus digital archiving.

Does the Case Study offer any indication of when it is safe for a library to switch to e-only delivery of e-journal content and so abandon the acquisition of paper copy? Some HE libraries, driven by user demand and the costs of paper storage, have already decided to collect substantial areas of the journal literature only in digital formats, abandoning paper copy. This is a brave action as they have had to do this lacking

[16] See http://www.kb.nl/dnp/e-depot/e-depot-en.html.
[17] See http://lockss.stanford.edu/. UCL has recently become a partner in the JISC-led LOCKSS consortium in the UK.
[18] See http://www.ithaka.org/e-archive/approach.htm.

- Any conclusive research into the comparative costs of paper and digital archiving
- An overview and analysis of digital lifecycles, including long-term preservation
- Reliable electronic archives for HE/FE and for HE researchers

The results of the UCL Case Study, and other Case Studies in the LIFE Project, show that there is still such uncertainty over the costs of digital archiving that it is too soon to abandon paper archiving for digital archiving. However, further work is needed on populating the LIFE formulae with real-life data before true cost models can be demonstrated. The models themselves are sound, but further Case Studies from copyright deposit and HE libraries are needed to get a firmer handle on long-term costs.

What does the UCL Case Study have to say about the drivers for moving to digital archiving, as outlined above and in the LIFE Project plan? Collection management policies in UK universities are driven by the needs of the research, teaching and learning communities they serve and at a pace which is suitable for the local academic environment. National, and copyright deposit, libraries are driven by different motives. They wish to collect the totality of the world's knowledge and to store it in perpetuity for the benefit of Society. This is a different motive. Academic libraries routinely dispose of materials no longer needed for research, learning and teaching. In a paper environment, copyright deposit libraries store for the longer term. The costs associated with digital archiving are such that it is not axiomatic that universities and copyright deposit libraries share the same agenda. The LIFE Case Studies indicate parallel activities which take place in both sectors, but as yet no clear agreement as to how work can be taken forward together in partnership. These are cultural as much as information management issues.

The UCL e-journals Case Study underlines the need to model the costs of long-term digital archiving on institutional budgets. The lifecycle approach is vindicated and the formulae adopted by the LIFE Project are robust, albeit with the minor addition of a T qualifier to the Aq element of the formula for HE. Certainly for HE, more rigorous data sampling over a longer timeframe with a larger number of partners in universities with different profiles would help the community to answer more of the Key Questions identified in the LIFE Project plan, and in more detail.

8. The Generic LIFE Preservation Model

8.1. Introduction and objectives

Identifying a cost for the preservation category of a digital object's lifecycle is particularly important as it has previously been identified as a recurring and potentially significant cost element[19]. There are a number of isolated examples of preservation action but very little costing information has been recorded. Few details are available of either the breakdown of what the process might involve or of the costs of each of those elements for the large scale preservation of digital collections.

The LIFE Project has therefore aimed to both identify and cost the different elements of digital preservation work which are likely to be required to support a digital repository containing an array of different types of digital materials.

Because of the lack of historical figures a strategy of estimation was employed. It should be noted that this is considered only the first step in developing an accurate and realistic costing model, one which will hopefully be refined as the experience of performing preservation is recorded over the coming years and in a future iteration of LIFE.

The key objectives of this work can be summarised as follows:

1. Making the first major step in defining and estimating the lifecycle cost of digital preservation activities.
2. Proposing a model for comment by the wider preservation community
3. Providing some rough cost estimates for "P" in the Lifecycle Model to enable the LIFE Case Studies to be compared and contrasted.
4. Attempting to identify the scale of preservation costs. Are they dramatically high as suggested previously by many in the preservation community or are they more achievable as suggested recently[20]?

8.2. Foundations for lifecycle preservation costing

The associated LIFE Research Review provides a detailed background to these preservation costing developments but the following works warrant particular mention.

The Nationaal Archief, Digitale Bewaring (2005) provides possibly the most detailed attempt to cost digital preservation activities and takes quite a different approach to

[19] See Cedars Project, Research Review.
[20] "Excuse Me... Some Digital Preservation Fallacies?", Rusbridge, C, http://www.ariadne.ac.uk/issue46/rusbridge/

that of LIFE. The Archief performed something more akin to a full accounting audit including a range of support and infrastructure costs which LIFE has deemed well outside the scope of a lifecycle costing approach. The Archief also focused on costing specific preservation strategies, again a contrasting approach to that of LIFE.

Oltmans, Kol (2005) offers a very useful first step in preservation activity costing using the lifecycle approach, which this work builds on. Again, the focus was made very much on comparing the costs of different preservation strategies. LIFE has taken a more generic approach in keeping strategy dependent factors to a minimum while delving deeper into a breakdown of the range of components of which the activities are formed. Kol and Oltmans include storage costs in their calculations, which LIFE has partitioned as a separate lifecycle category.

Dürr and Meer (2001) provide some interesting costings of a small scale archive, and include costs for storage and preservation activity.

8.3. Developing the model

Given the lack of hard evidence on which to base the model, a number of review processes were used in an attempt to refine the estimations used as much as possible.

1. Following the development of an initial draft of the model, a component estimation review was conducted with two members of the BL eIS Architecture team. Each key component of the model was graphed out on the basis of what the reviewers thought the trends for that component would be. This was then compared to what the model actually predicted, and the comparisons were discussed, analysed and where necessary changes were made to the model. On the whole, the reviewer's projections matched the model for the 20 year timescale quite closely, but some minor changes were made.
2. The model was tested using the file format data from the VDEP and Web Archiving Case Studies, allowing the output of the model to be considered. These two Case Studies provided a good range of data and quite contrasting results. The Project team reviewed the data and refinements were made to the model inputs.
3. A further review of the output data was made by the eIS Architecture team. A number of changes were made to the model where obvious weaknesses were identified (for example the initially linear modelling of migration costs was enhanced to a rational model to represent economies of scale).

Limited resources and tight timescales for Project deadlines did not allow time for external review of the model. However LIFE hopes that following publication of this work, the wider preservation community can now comment and preservation

costing can be taken forward in further work in the near future. Suggestions for further development can be found below.

8.4. Modelling digital preservation costs

8.4.1. Key elements

A range of key factors were identified as being significant in the preservation of digital objects. The following list[21] is certainly not exhaustive, but provides a starting point for modelling a complex process for which little practical costing evidence exists:

- Frequency of action – how often preservation action needs to be taken
- Technology watch – identifying the points at which preservation actions need to occur
- Availability of tools – how often tools are available and therefore a new tool does not need to be developed specifically for the purpose
- Complexity of file formats – how the file format itself affects the cost of a preservation action
- Updating metadata – recording the crucial metadata which describes preservation tools and actions
- Cost of tools – depending on the availability, the cost of acquiring or developing a rendering solution
- Preservation strategies – how the model addresses the use of different approaches to preserving digital objects
- Preservation action – the activities involved in performing preservation actions
- Quality assurance - checking the accuracy and effectiveness of a preservation action

In the explanations below, "t" is the time and "n" is the number of objects of a particular file format at the time of preservation. The model is considered to be valid for estimating costs from t=0 to t=20 years.

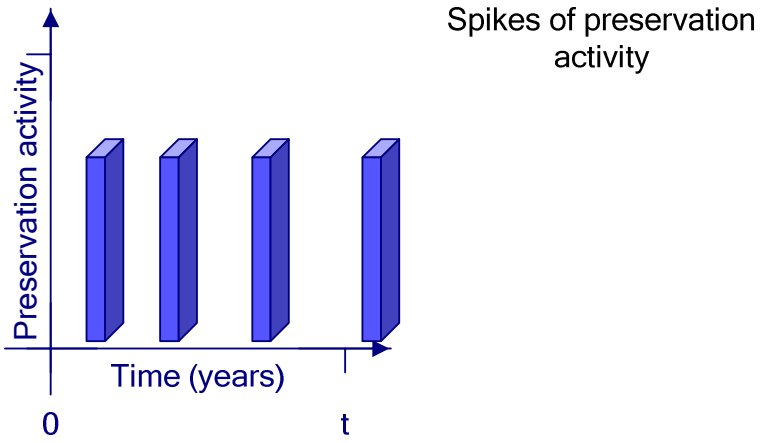
8.4.2. Frequency of action

As Rusbridge indicates[22] it is perhaps very rare for file formats to become completely inaccessible. Despite this, there is certainly a transition point at which file formats stop becoming readily viewable at the click of a mouse on a modern computer and become harder to access. When this occurs, a digital repository must put in place a new access mechanism or "rendering tool" to enable users to view retrieved objects of this format. Metadata describing the method of rendering

[21] Note that the order of this list has no particular significance other than facilitating the explanation of the elements within the model.
[22] "Excuse Me... Some Digital Preservation Fallacies?", Rusbridge, C, http://www.ariadne.ac.uk/issue46/rusbridge/

needs to be changed, tools need to be acquired or developed and the objects themselves may need to be migrated to a different format.

This results in spikes of preservation activity at recurring intervals over time.



Estimating the time between spikes in the future is difficult. Considering historical evidence and experience with formats over the past 25 years has led the Project team to propose a base frequency of once every 8 years[23].

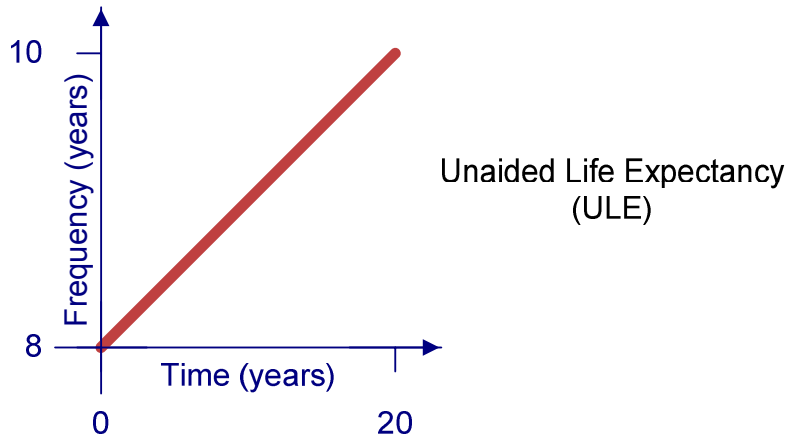
BLE is the **Base life expectancy** and is a model input which can be easily modified in the associated spreadsheet when applying the model.

BLE = 8

The consensus among experts is that file formats are maturing and becoming more long-lived.[24] We expect the base figure to increase over time. The spikes of preservation activity then become less frequent over time, as shown in the diagram above.

LIFE has therefore modelled the frequency of preservation action as increasing by 1 year for every 10 years that pass from the current time.

[23] Dürr and Meer (2001) suggested around 5 years.
[24] Evidence of file format standardisation activities by key commercial developers like Microsoft and Adobe appears to back this up.



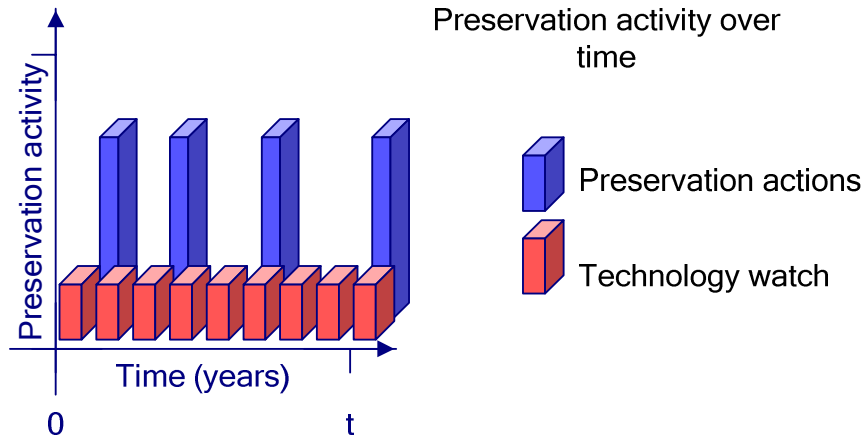
The model terms this frequency the **Unaided Life Expectancy of a Format** or **ULE**.

$$ULE = BLE + 0.1*t$$

8.4.3. Technology Watch

Technology Watch is the process of monitoring a particular file format, the tools that render that file format and the software and hardware infrastructure those tools run on in order to provide an alert when preservation action needs to be taken in order to ensure the continued use of data in that format. This process involves monitoring the technology involved and recording and updating metadata about that technology. In many cases, this will involve only minor updates to Representation Information metadata. The additional costs of addressing issues when a file format becomes obsolete are included elsewhere (UME).

The model assumes that Technology Watch will be performed for each file format once per year. This results in preservation activity as shown below.



We have conservatively estimated this as a week of time per format by a metadata officer.

The model terms this the **Technology Watch per Format** or **TEW** and is estimated as the work of one metadata officer for one week at an annual salary of £30k[25]. **TEW** is a model input which can be easily modified in the associated spreadsheet when applying the model.

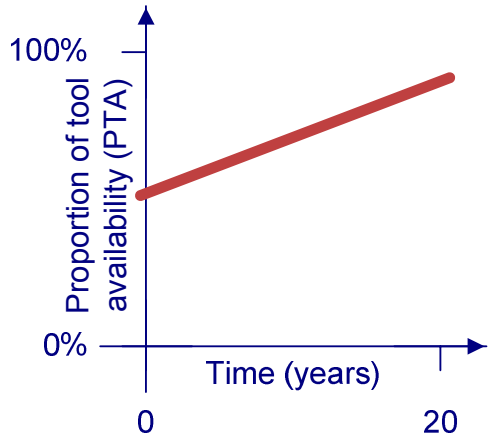
TEW = *£625*

8.4.4. Availability of tools

When a spike of preservation action occurs a range of work has to be undertaken. A key element of this work depends on whether an appropriate tool is available (either as open source or a commercial purchase) or a tool has to be developed which is potentially much more expensive.

LIFE proposes that, on average, tools will be available for a particular file format about 50% of the time, at t=0 years.

As digital preservation funding increases over time, the availability of preservation tools will increase. The model predicts that this availability will reach about 90% in the next 20 years.



Predicting this trend is very difficult, but it is clear that preservation funding is on the increase.

EU funding in this field has rapidly increased from virtually nothing to 30-40 million euros in the current FP6. It may well be over double this for FP7. JISC funding in

[25] This represents an estimated indicative salary, including on costs, and based in London.

the UK is about £6 million for the current 3 year Digital Preservation and Records Management Programme[26], and has increased dramatically from previous years.

More of this digital preservation funding appears to be going into practical tool and service development. Examples include the EU-funded PLANETS Project which is developing an array of preservation tools, the KB/NA emulation Project which is developing a modular x86 emulator,[27] and the Mellon-funded Global Digital Format Registry Project[28]. It is also possible that open source enthusiast developed tools are on the increase, although this is difficult to quantify. Commercial tools may play a significant role in the future, but this area is also difficult to predict.

The model terms this element the **Proportion of Tool Availability** or **PTA**. It calculates average tool availability for the period 0 to t.

PTA = **STA**(1-t/20) + **ETA**(t/20)

STA is the Starting Proportion of Tool Availability. **STA** = 0.5

ETA is the Ending Proportion of Tool Availability. **ETA** = 0.9

Both **STA** and **ETA** are model inputs which can be easily modified in the associated spreadsheet when applying the model.

8.4.5. Complexity of file formats

Various aspects of a file format impact on the costs of preserving objects comprising format. These include: size, complexity, whether it is open, standardised or proprietary, and so on. For the purposes of this model, LIFE proposes a single indicator which will provide a basic representation for some of these factors. This will be utilised in different aspects of the model (see below).

The **File Format Complexity** or **FCX** is a linear scale from 0 to 1. A number of categories simplify the allocation process as shown below:

Category	FCX	Examples
Simple	0.1	ASCII, Unicode
Bitmap	0.2	JPEG, GIF
Mark-up	0.3	XML, HTML
Vector	0.4	EMF, Draw
Multimedia	0.6	MPEG3, WAV
Document	0.8	Word, PDF
Complex	1	Oracle database dump

[26] http://www.jisc.ac.uk/index.cfm?name=programme_preservation
[27] http://www.digitaleduurzaamheid.nl/ will have a link to the emulation project's source forge site shortly.
[28] http://hul.harvard.edu/gdfr/

FCX categories are model inputs which can easily be modified in the associated spreadsheet when applying the model.

8.4.6. Updating Metadata

When a preservation spike occurs for a specific file format a new rendering solution is selected, acquired, and representation information describing it is updated. The process of selection and recording might involve:

- Research into possible preservation solutions
- Selection of an appropriate solution
- Review and approval by a panel of experts
- Recording of Representation Information describing the new rendering solution

It is unclear exactly how much work this would involve and how much review and revision of the choices there would need to be. An indication can be made from the work of the Florida Centre for Library Automation to develop preservation action plans[29] for a number of common file formats. On average, each file format plan took 10 days to develop. This process will be roughly estimated as 2 weeks work for a metadata officer at an annual salary of £30k[30].

For the purposes of the model this process has been termed **Update Metadata** or **UME**, but as the text above indicates this may develop into a more involved selection and review process.

UME = 2 metadata officer weeks @ £30k annual salary = £1250
UME = £1250

UME is a model input which can easily be modified in the associated spreadsheet when applying the model.

8.4.7. Cost of tools

The availability of preservation tools has already been discussed above, and indicates the proportions in which the following possibilities occur:

1. Develop a new tool
2. Acquire a commercial or open source tool

By combining the calculated proportions of each of these eventualities with estimated costs for those eventualities we can arrive at a single average cost for a new tool. LIFE terms this the **Cost of a new rendering solution** or **CRS**.

The cost of an available rendering solution is proposed as a constant, which has been termed **Cost of available tool** or **COA**. The **COA** could potentially represent

[29] FCLA, http://www.fcla.edu/digitalArchive/daInfo.htm
[30] This represents an estimated indicative salary, including on costs, and based in London.

the cost of implementing and integrating a tool into a repository work flow as well as the cost of the tool itself. The tool could be open source and effectively free to acquire, or it may be commercial and available at a price. LIFE proposes a nominal cost of £1500 for the **COA**.

COA = £1500

COA is a model input which can easily be modified in the associated spreadsheet when applying the model.

Combining this with the calculated proportion for which tools (on average) are available:

Working available cost = **PTA** * **COA**

The cost of developing a tool may well be far higher. Over time this cost may increase as file format structures get bigger and more complex. As experience and infrastructure in developing digital preservation tools increases, the costs may decrease. Predicting these trends is very difficult. For the purposes of this model it is assumed that this cost will remain constant over time.

LIFE proposes that a starting point for this cost as an input to the model would be a development cost defined loosely as 24 programmer months at an annual salary of £30k[31]. LIFE defines this as the **Tool development cost** or **TDC**.

TDC = 60000

The complexity of the file format is considered to be a major factor in the size of the tool cost. The **FCX** (see above) is therefore used to scale

Working development cost = **TDC** * **FCX**

Combining this with the calculated proportion for which tools (on average) are available:

Working development cost = (1-**PTA**) * **TDC** * **FCX**

Finally, these working costs are combined to calculate an average **Cost of new rendering solution** for the period modelled:

CRS = (1-**PTA**) * **TDC** * **FCX** + **PTA** * **COA**

[31] This represents an estimated indicative salary, including on costs, and based in London.

8.4.8. Preservation strategies

LIFE favours the design of a generic model rather than a series of preservation strategy-dependent models for a number of reasons:

- Specific detailed models are difficult to cost without evidence of real costing data
- The cost, selection and proposed use of different preservation strategies can be a highly contentious subject. LIFE is keen for the time being to focus discussion on costings rather than the debate regarding the relative merits of different preservation strategies

However, it became clear that some elements of the costing model required preservation strategies to be taken into account, at the very least, in a basic way.

Without getting drawn into the preservation strategies debate, LIFE suggests that a range of preservation strategies will be required to preserve a cross section of different kinds of materials.

LIFE therefore proposes that the Generic Model represents a range of different strategies. These can loosely be considered as Normalisation/Migration, Emulation and Migration on Request. Life proposes that one very significant impact of preservation strategy should be taken into account. Normalisation/Migration is considered a special case with regard to how preservation action is costed because it is dependent on the number of objects being preserved and occurs when the object is first ingested to a digital repository.

For simplification, LIFE suggests that the Normalisation/Migration strategy will occur 40% of the time and terms this the **Proportion of normalisation** or **PON**. The remaining 60% of the time, alternative strategies like emulation or migration on request occur with no object dependent costs in preservation action.

PON = 0.4

PON is a model input which can be easily modified in the associated spreadsheet when applying the model.

8.4.9. Preservation action

Preservation action is the process of performing some kind of preservation activity on a number of objects. This might include:

- Setting up a preservation process
- Performing migration on a batch of files
- Recording metadata about the preservation action
- Re-ingest of migrated files into the repository

Note that QA is costed separately (see below)

A simple way of modelling this cost is to use a proportional scale based on the number of files (n) being preserved:

Working cost of preservation = n * PCP

PCP is the **Per object cost of preservation** and is a working constant.

Experimentation conducted as part of the VDEP Case Study attempted to produce an informed estimate of the constant per object cost. For about 2000 objects this was estimated to be £0.22.

PCP = 0.22

This produces the following results for a range of values for n (note that the n=2000 value used in the VDEP trial is included here):

Number of files	1	10	100	1000	2000	10000	100000	1000000
PCP	£0.22	£0.22	£0.22	£0.22	**£0.22**	£0.22	£0.22	£0.22
Cost for n objects	£0.22	£2.20	£22.00	£220.00	**£440.00**	£2,200.00	£22,000.00	£220,000.00

There is a problem with this approach. The cost of preserving a small number of files is very small, despite the overhead of the process setup experienced in the VDEP experimentation. As a consequence, no efficiency is gained through economies of scale[32], so very large numbers of objects are very expensive to preserve.

A slightly more realistic approach is to include a setup cost for the preservation activity and utilise a formula that tends towards a defined high volume cost per object. This models a higher total cost per object for small numbers of objects and models a lower cost per object for high numbers of objects where efficiencies can be made.

LIFE proposes the **Setup cost of migration** at around £340.

SCM = £340

[32] A key point made by Oltmans and Kol (2005), who suggested a per object migration cost of $0.1 per object (equating roughly to £0.05 HVM proposed by LIFE)

Life proposes a **High volume migration cost per object** or **HVM** at £0.05. This is the value that the cost will tend to for a high number of objects.

HVM = £0.05

Working cost of preservation = **SCM** + n * **HVM**

Number of files	1	10	100	1000	2000	10000	100000	1000000
cost per object with setup cost	£340.05	£34.05	£3.45	£0.39	**£0.22**	£0.084	£0.053	£0.050
total cost for n objects with setup cost	£340.05	£340.5	£345.0	£390.0	**£440.0**	£840.0	£5,340.0	£50,340

As shown above this produces a reasonably realistic curve and low end cost, and also replicates the VDEP experimentation cost for 2000 objects as £440.

LIFE terms this the cost of **Performing preservation action** or **PPA**, and models this cost as:

PPA = **PON** * (**SCM** + n * **HVM)**

Note that **PON** is added as a multiplier as this cost is only present for the **Proportion of normalisation** or migration performed (see Preservation strategies above).

8.4.10. QAA

After performing a preservation action, a process of quality assurance is required to ensure the action has met with a required level of accuracy. QAA is likely to involve a visual and perhaps automated comparison of the original and the preserved objects. The sample of objects tested could vary tremendously depending on the requirements and value of the collection being preserved. Low volume, high value resources might require every object to pass a QA test. High volume, low value collections may only need a small sample of objects to be checked. LIFE therefore proposes a middle ground based on experimentation. An input to the model allows this value to be adjusted appropriately.

Experimentation conducted as part of the VDEP Case Study attempted to produce an informed estimate of the constant per object cost for this process. For about 2000 objects this was estimated to be £0.17. This is scaled by the format complexity.

LIFE terms this the **Base cost of testing a preservation action per object** or **BCT**.

BCT = 0.17.

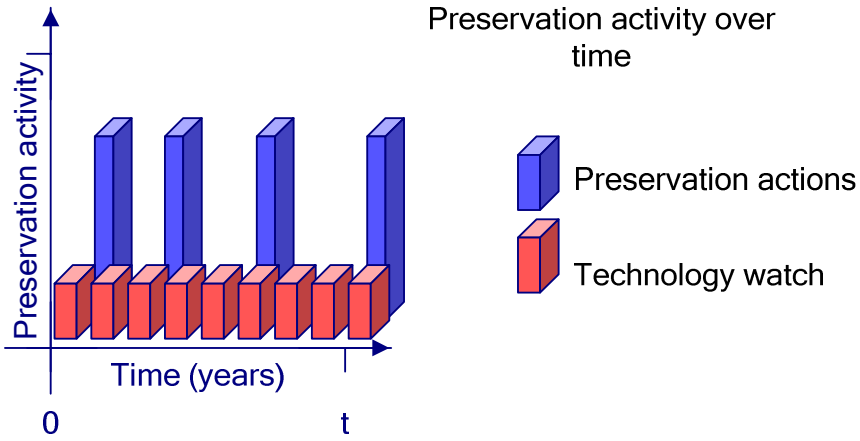
BCT is a model input which can be easily modified in the associated spreadsheet when applying the model.

The overall cost is defined as the **Quality assurance** or **QAA**:

QAA = n * **BCT** * **FCX**

8.5. Combining the elements – the Preservation Model

The preservation cost for a particular file format from time=0 to time=t consists of both a regular Technology watch cost and less frequent spikes of preservation activity when action is required to ensure continued access to the format.



The **Technology watch** cost consists of annual spikes of activity, defined by **TEW**. The overall cost is therefore:

Technology watch cost = t * **TEW**

The **Overall preservation action** cost consists of spikes of activity occurring a number of times as defined by the **Preservation frequency**. Dividing the length of time the calculation is required for by the **ULE** gives the number of actions required in that time period. Normalisation occurs at the start of the time period and so **PON** is added to this result.

Preservation frequency = **t** / **ULE** + **PON**

The **Overall preservation action** cost is a summation of the key areas of activity present in each spike of preservation action. This includes the **Cost of new**

rendering solution, the **Update Metadata** cost, the **Performing preservation action** cost and the **Quality assurance** cost.

Overall preservation action = **CRS** + **UME** + **PPA** + **QAA**

Combining these three key parts, provides the complete preservation cost:

Preservation = Technology watch + Preservation frequency * Overall preservation action

Preservation = **t * TEW** + **(t / ULE** + **PON) * (CRS** + **UME** + **PPA** + **QAA)**

8.5.1. A useful breakdown of the Preservation cost

A useful breakdown of the preservation costs can be made by rearranging the formula as follows:

Technology watch	= **t * TEW**
Preservation tool cost	= **(t / ULE** + **PON) * CRS**
Preservation metadata	= **(t / ULE** + **PON) * UME**
Preservation action	= **(t / ULE** + **PON) * PPA**
Quality assurance	= **(t / ULE** + **PON) * QAA**

An example taken from the Web Archiving Case Study produces the following results for a 20 year period:

File Format (MIME)	Format Complexity (FCX)	Estimated objects per year	*Total tech watch*	*Total tool cost*	*Total UME*	*Total PPA*	*Total QAA*	Total
image/png	0.2	8474	12,500.00	6,120.00	3,000.00	733.14	691.45	**£23,045**

8.6. Summary of the Generic LIFE Preservation Model

The preservation cost for n number of objects of the same file format[33], over a period of t years beginning at the present time, is:

Preservation = t * TEW + (t / ULE + PON) * (CRS + UME + PPA + QAA)

Expansion of calculated components:

- ULE – Unaided Life Expectancy of a Format = BLE + 0.1*t
- CRS – Cost of new rendering solution = (1 - PTA) * TDC * FCX + PTA * COA
- PPA – Performing preservation action = PON * (SCM + n * HVM)
- QAA – Quality Assurance = n * BCT * FCX
- PTA – Proportion of Tool Availability = STA(1-t/20)+ETA(t/20)

Expansion of scaling components:

- PON – Proportion of normalisation = 0.4
- FCX - Format complexity (e.g. JPEG = 0.2, WMF = 0.4, PDF = 0.6, Word = 0.8)

Expansion of cost component inputs:

- HVM – High volume migration cost per object = £0.05
- BCT – Base cost of testing a preservation action per object = £0.17
- UME – Update Metadata = 2 metadata officer weeks @ £30k annual salary = £1250
- TDC – Tool development cost = 12 programmer months @ £30k annual salary - £60000
- COA – Cost of available tool = £1500
- TEW - Technology Watch = 1 metadata officer week @ £30k annual salary = £625
- BLE - Base life expectancy = 8 (years)
- STA – Starting tool availability = 0.5
- ETA – Ending tool availability = 0.9
- SCM – Setup cost of migration = £340

[33] For the purposes of this model, a "format" is considered to be a distinct version of a file format that requires specific consideration with regards to preservation.

8.7. Evaluation of the model

8.7.1. The broader picture

As a process, both developing the model and applying it to a real collection has been a valuable experience for the LIFE team, and the same will hopefully be found by those considering the model. Applying the lifecycle approach to preservation costing provides a new perspective into the range, scope and design of preservation activities that will be necessary to ensure long term digital preservation.

On one hand it is very easy to argue that the headline cost outputs from the model are based to such a large degree on estimation that they are virtually useless. At the same time, consideration of the component costs can provide some very useful results. Even if the model's results are inaccurate, it can still be used to drive a strategic approach to preservation. Where do the key component costs lie? What factors control these costs? Where can savings are made? Which elements of the preservation lifecycle can be shared? On which processes should automation efforts be concentrated? Analysis of implementations of the model provides at least some steer on all of these questions.

8.7.2. Evaluation against key objectives

Make the first major step in defining and estimating the lifecycle cost of digital preservation activities.

The model has fulfilled this objective well. It significantly advances work in this area, and provides a good basis for further development.

Propose a model for comment by the wider preservation community

Again the model achieves this objective well. A key issue will be how this work is taken forward and it is hoped that discussions at the LIFE Project Conference will provide a useful steer on this.

Provide some rough cost estimates for "P" in the Lifecycle Model to enable the LIFE Case Studies to be compared and contrasted.

Useful comparisons can be made between the Case Studies using the model despite the unreliability of the headline costed figures. The model highlights the need for dramatically different approaches to the preservation of the VDEP and Web materials.

Attempt to identify the scale of preservation costs. Are they dramatically high as suggested previously by many in the preservation community or are they more achievable as suggested recently?

The model does not appear to offer a clear indication one way or the other[34] but does reinforce the clear need for collaboration amongst leading institutions involved in digital preservation activities, sharing of digital preservation developments, and cost effective digital preservation service models. Many of the key costs identified by the model can certainly be shared by the preservation community.

The headline figures for preserving each of the Case Study collections appear quite high but the component costs for a particular file format feel reasonably realistic. This perhaps suggests that the total cost of preservation investment will indeed be well beyond the means of any individual institution (prompting the views of the pessimists), but the possibilities for sharing, automation and reuse of that investment will result in far more realistic costs in the medium to long term (the more enlightened or perhaps optimistic view!). This partly explains the high short term headline costs and the far lower 20 year headline costs produced by the model.

8.8. Further work

A range of suggestions are made for further development of the model:

- The most obvious suggestion would be to refine the model using real life cost data. Follow up work could engage more actively in recording cost/effort data for preservation activity (perhaps as part of collaborations with new or ongoing preservation projects), or at the very least in promoting a strategy for others to record cost and publish cost data for the greater good. LIFE notes that there is often a reluctance to record this kind of data when the focus of the activities is often on very new or experimental work.
- The format complexity is currently represented as a linear scale but this does not sufficiently capture the range of complexity that might be encountered. In reality a database may be several magnitudes more complex than a simple document or bitmap graphic. A development of the model would greatly benefit from a comprehensive revision of the FCX, which might include:
    - A new method of representing the complexity scale.
    - More detailed examination of the format categories and the location of these categories on the complexity scale.
    - Consideration of other factors, like proprietary versus open source.
    - Associating frequency of required preservation action with category of file format
- Preservation actions like normalisation will often result in a repository having to deal with fewer file formats as preservation activities progress. This is difficult but perhaps crucial to model. Some preservation tools can provide solutions for more than one format. Again, modelling this will be difficult but ultimately very useful to cost.

[34] Although this is very much a judgement call which external comment will no doubt answer more clearly following publication of this work!

- A significant amount of the effort involved in a migration based preservation action is likely to be associated with re-ingest of a the resulting migrated object to the repository. Further work should explore the integration of Ingest costing with the Preservation model in order to calculate it more effectively. Evidence from real re-ingesting activity will also be useful to inform the costing.
- Developing the model to represent the nuances of different preservation strategies in detail would be a useful exercise, but only if it can be backed up with hard data. The relative merits of different strategies can quickly become a controversial issue, which is why the current model remains largely strategy-neutral.
- Examine alternative approaches to modelling the preservation cost. The use of average costings across the timescale is somewhat crude and is a consequence of the explorative and iterative way in which the model was developed. An annual summation approach would be a useful starting point for further work.
- The Cedars and CAMiLEON Projects based their research and development work on the key principle that much could be learnt about future trends and preservation work that would need to be performed, by examining materials and other evidence from recent history. Can comprehensive research into past trends (and previous preservation action) provide valuable evidence for the Preservation Model? This might include:
    - Spending on digital preservation activities
    - File format obsolescence
    - Longevity/lifetime of commercial tools

9. LIFE Project findings

The LIFE Project Case Studies have proven to be highly effective in highlighting both the types of issues that can be encountered in a digital collection and the ways in which a lifecycle methodology can be utilised to capture and apply a cost to these problems. This combination of real data, and a framework within which to apply them, has simplified many of the perceived areas of complexity within the digital preservation arena. In so doing, the following findings have been identified and extracted from the previous Case Studies. The Project has successfully identified some key themes and some consistent messages which are brought together from these findings in the final conclusion. The specific Project findings from LIFE are listed below.

9.1. VDEP Case Study

9.1.1. Project findings

1. The cost of storing and preserving an e-monograph varies by file size. The LIFE Project estimates a total cost of £14-£20 per e-mono in year one, rising to closer to £40-£50 by year five. For an e-serial, this range is £23 in year one dropping to £8.55 per issue by year 5. (see section 5.6 for further detail)

Example

Example	**Yr1**	**Yr5**	**Yr10**
Instructors CD-Rom	14.53	22.51	29.79
The numbers crew	29.81	68.39	86.32
AGI Geographic	14.14	21.33	28.26
Average cost for e-monographs	**19.49**	**37.41**	**48.12**

1. Both serials and monographs bring their own cost problems. Monographs can vary from 1kb to 1000mb and serials also vary considerably in both frequency and size. This makes selection a key area for policy development.
2. The average preservation cost per entity in VDEP is £0.089p based on a 5-year cycle.
3. Preservation costs are projected to go down over time, not up, for this collection using the LIFE Preservation Model.

4. There are, as yet, no obsolete file formats within VDEP and indeed LIFE struggled to find any formats at risk.
5. Project examples for the cost of bit-stream storage for e-monos or e-serials varied from £3.00 to £11,500 for a 20-year life.
6. It cost £80.85 per gb in the first year to store the whole VDEP collection in year 1. By year 5 this has dropped to £16.17 per gb.
7. If a DOM approach had been taken, the costs would have been £27.45 in year 1 and £5.48 in year 5.
8. 230,000 separate files have been deposited over 5 years (172,484 objects).
9. DigiTool 2.3 is not adequate or scaleable for a large-scale repository due to its lack of automated functions.
10. Zipping is unsuitable for any large scale archive or true digital preservation repository.
11. Ingest is currently a very manual process and in its present form incurs a high lifecycle cost.
12. Metadata is still largely undefined and manual. It can contribute to as much as 50% of Ingest costs and up to the same again over the total lifecycle costs (excluding storage) due to lack of automation and extraction.
13. Creating new catalogue records for VDEP is the same cost as creating current records for analogue items, and both are done on Aleph.
14. Access falls outside the scope of the VDEP collection.
15. The average item size is 7.75 mb. The average object size is therefore 10.30 mb.
16. Migration is feasible for the small part of this collection that was tested. More work is required with the correct tools to establish whether this is applicable to a wider range of formats. Other strategies have not been ruled out and must be tested alongside this work.
17. 1.34 items constitute a bibliographic record (or object).

9.1.2. Strategic findings

18. Large scale investment at the Ingest point to automate metadata would vastly reduce processing costs.
19. Standard Metadata schema development is a crucial for a digital repository. A national standard must be developed.
20. Anything that falls outside this standard workflow will need to be dealt with by Special Collections teams.
21. It is far more challenging and expensive to apply preservation metadata to hand-held resources due to their complex file structure and relationships. Future work must focus on this area and report findings.
22. The smaller the file size, the greater the cost (comparatively) to collection areas. The higher the file size, the greater the cost (reality) to store.
23. Access is largely going to be determined by legislative changes. Without clear guidance on who has the legal right to access what, digital rights management becomes grey.

24. The reduction of storage costs over time balances exactly with the ongoing increase in serial issues. In other words, the larger the archive becomes in mb, the storage costs over time remain constant (assuming storage space is infinite and no inflation is added). E.g. Year 1: 4 serials = £10 Year 5: 20 serials = £10.
25. The pragmatic DOM approach delivers a strong cost reduction at face value. Redundancy and reliability issues need further work to estimate value.
26. The LIFE study clearly shows that Architecture standardization is essential. However, the study was not able to make a conclusion as to the relative cost effectiveness of a national/copyright deposit library vis-à-vis an HE library taking on digital preservation as a national responsibility. As with paper archiving, the identification of just one body as having responsibility for digital archiving is a risk, as there is only one single point of failure. It should be noted, however, that the BL is further ahead than HE in terms of adopting and costing a lifecycle approach to the long-term management of digital assets. The issue of cost-effectiveness and responsibilities for digital archiving need to be investigated more deeply in a further phase of the LIFE Project.
27. Selection and Storage are inseparable. Close monitoring of each and amendments to policy must be conducted for an electronic collection.
28. Preservation is possible on a large scale. Costs are projected to go down not up using the LIFE model. Work needs to be continued in this area to test the validity of this finding on a large scale.
29. Tool development to cover all major file formats must be created. Strategically this is important and widespread collaboration should be agreed. Repositories across the UK should be targeted to take part in a national requirements project.
30. The lifecycle methodology and approach is robust and worth writing into strategic plans. (BL are adopting this as a standard approach to long-term curation). The Generic LIFE Preservation Model should be rigorously tested alongside this work.
31. In a five year collection of electronic data, NO obsolete file formats were discovered. Extensive testing and research indicate that everything within the VDEP collection can be salvaged if required. This is a key strategic finding for future work.
32. A national approach to technology watch and representation information is a required digital preservation resource for standardisation across institutions.
33. Automated metadata extraction procedures and negotiations to include preservation metadata where possible within the Legal deposit framework are required in order to protect both the long-term access and preservation of the National collection and to reduce the potential long term cost of preservation.

34. Future project work comparing analogue storage and preservation costs to digital lifecycle costs is strongly recommended by the LIFE Project for the HE/FE and Library sectors and domains.

9.2. UCL e-journals Case Study

9.2.1. Project findings

1. In terms of their management of commercial e-journal content, HE institutions currently do not preserve digital content for the long-term. The lessons from the LIFE Project are that considerable development of the workflow processes would have to be undertaken to embed digital preservation into the institutional management of e-content
2. The Library which forms the subject of the study is extremely cost-effective in terms of the way it manages e-journal content. Over a 10-year period, the costs projected in the LIFE formula per title are

LIFE element	Cost for T=10 years	% of total cost
Aq		
Purchase	£5105.71	97.40
Staff	£53.36	1.02
I	£0	0
M	£3.97	0.08
Ac		
Staff Member A	£44.69	0.85
Staff Member B	£34.84	0.66
S	£0	0
P	£0	0
TOTAL	**£5242.57**	**100%**

3. Only 2.61% of the lifecycle costs over 10 years can be attributed to staff costs in the UCL Case Study.
4. Universities deliver e-journal content which has a financial value of over thirty seven times the activity costs associated with making it accessible
5. The HE library which was the subject of study needs to undertake further work to embed activity costings into its procedures
6. A further Case Study, over a longer timeframe with more detailed costings, would help to populate the LIFE formulae with more robust data.

9.2.2. Strategic findings

7. The lifecycle approach to the costing of e-journal acquisition, access and archiving is robust
8. UCL's workflows, typical of a service-driven HE library, lend themselves to the lifecycle approach

9. UCL would need to invest considerably at the technical, financial and workflow levels to embed digital archiving procedures into its current operations
10. The conclusions presented elsewhere in this Report show that the British Library is further advanced than HE in terms of lifecycle costings and the construction of a digital repository for the long-term storage of digital content. The present Case Study suggests, however, that digital archiving at a local level is an issue which universities should consider seriously.
11. Libraries, which have already abandoned paper acquisition in favour of e-only delivery for journals, have done so in the knowledge that no reliable digital archives exists in the UK, which is a brave decision
12. A second phase of the LIFE Project would enable UCL and other universities in the UK, with the British Library, to populate the LIFE formulae with robust data over a longer timeframe, which would help the community to identify a way forward for digital archiving at a national level

9.3. Web Archiving Case Study

9.3.1. Project findings

1. Further case studies focusing on web materials in more established Web Archiving activities would be useful in providing more evidence for the findings. The costs of domain-wide Web Archiving also need to be explored.
2. A % break down table for the Web Archiving lifecycle cost reveals the following:

LIFE model category	Cost per title[35] after 10 years	% of total cost
Acquisition (Aq)	£934.09	13.77%
Ingest (I)	£1,114.51	16.16%
Metadata (M)	£4.25	0.12%
Access (Ac)	£29.57	0.45%
Storage (S)	£539.08	7.82%
Preservation (P)	£4,254.96	61.69%
Total	**£6,876.46**	**100.00%**

3. The average lifecycle cost for archiving a new web site title is £21.28; and, for archiving a single instance of that title, the average cost is £130.30 for one year.
4. The complete lifecycle cost of archiving a title at the average rate of just over 5 instances per year for 20 years is £13,732.

[35] Cost per title includes the costs of an average of just over 5 instances per year.

9.3.2. Strategic findings

5. Although many of the Web Archiving processes differ from what might be considered the norm, the LIFE Project Lifecycle Model proved more than suitable for calculating and comparing the costs of Web Archiving activity.
6. The Generic LIFE Preservation model indicates that the cost of preserving web materials will be high, particularly in the short term. Preservation represents approximately 55% of the complete lifecycle costs. Automated processes and on-demand techniques, whose costs are not volume-dependent, will be essential in meeting the preservation challenge. Investment in tool development will provide an extremely important foundation.
7. The current Web Archiving activities are in their infancy in terms of scale, but also in terms of the capture of content. Collection and recording of Metadata, the execution of characterisation of the content for the purposes of preservation, and the capture of the context of the selected sites are key areas for development. The costs of these operations need to be investigated.
8. Greater efficiencies, and the introduction of more automated processes, will reduce Web Archiving costs considerably, but unavoidable manual effort is likely to leave Ingest at a relatively high level for the medium term. The likely introduction of Legal Deposit legislation covering web materials will dramatically cut the cost of the IPR portion of the Acquisition costs.
9. Costing activities are themselves at a very immature stage of development. The models, techniques and outcomes of the LIFE Project and other work will need to be developed and refined in order to provide useful results for preservation planning. Recording and utilising real life cost and activity data (particularly in the areas of preservation and access) will be crucial in achieving this.
10. Support, management, administration and many other costs have not been included in the LIFE Project's lifecycle approach, but they are considerable. It is not clear from the Case Studies whether this approach to the scope of costing activity is useful or not.
11. The process of identifying the different elements of a digital object's lifecycle, and then costing those elements, provided a very useful insight and approach to the challenges of digital preservation, beyond the obvious outputs of costing data useful for strategic planning.
12. Performing a lifecycle-based costing exercise may provide some negative outcomes, particularly if sensitive cost/activity data is revealed to the outside world. LIFE has been somewhat courageous in exposing this kind of internal information and hopes that the benefits of this approach will be seen to outweigh the possible negative aspects.

10. Conclusions

The LIFE Project has established that a lifecycle approach to cost is both applicable and useful for a range of digital collections. The three Case Studies, which are vastly different in both content and workflow, have as expected returned three very different outcomes. However the variations in cost and workflow have been successfully captured within the lifecycle model and the associated Generic LIFE Preservation Model.

The VDEP's costs are strongly weighted in the areas of metadata and storage. This contrasts with the high acquisition and access costs for e-journals and Web Archiving's preservation costs. However, the LIFE model is able to capture all of these distinct trends and gives us the belief that it can be used to capture a snapshot of any digital collection at any point in time. This positions the model well for future project work.

All exemplars picked up on the fact that tool development for digital preservation is a high priority and means that the model can only go so far without help. There are significant costs to be saved, in areas such as ingest and metadata, if the correct tools are able to be developed for all aspects of the lifecycle of digital collections.

As reported in the Web Archiving findings and in the UCL e-journals Case Study, costing activities are themselves at a very immature stage of development. The models, techniques and outcomes of the LIFE Project and other work will need to be developed and refined in order to provide useful results for preservation planning. Recording and utilising real life cost and activity data (particularly in the areas of preservation and access) will be crucial in achieving this.

A second phase of the LIFE Project is recommended, as this would enable UCL and other universities in the UK, with the British Library, to populate the LIFE formulae with robust data over a longer timeframe, which would help the community to identify a way forward for digital archiving at a national level. Future project work comparing analogue storage and preservation costs to digital lifecycle costs is also strongly recommended to provide better information to guide selection policy in a hybrid analogue/digital collection.

Digital preservation costs are predicted to go down over time using the LIFE model and VDEP. It is important to understand that this conclusion relates to the unit cost only. The whole repository cost of including preservation will of course go up as more content is added, but technology advances, tool development and experience indicate that the unit cost to preserve an item will reduce over time.

This concept needs to be applied to more collections in order to validate it. Web Archiving predicts that this too is the case, although the starting point for

preservation within Web Archiving is very high. Technology advances over time and the reducing costs of tool development should also contribute to this outcome.

There are (as yet) no obsolete file formats discovered within the LIFE Project, and this finding has been consistent across all three Case Studies. The majority of this material is however quite recent. All three studies showed some level of proprietary formats but only the UCL e-journals study is the found formats that we had little knowledge of (gff, .briggsae and. elegans, within its PLoS collection) that caused concern. However, these file formats were identified not as obsolete, but in need of better description. Developing computer technology means that without good information older formats could become inaccessible and so must form a part of any preservation strategy. Further work in this area to test this finding is strongly encouraged, and a robust technology watch is required to guarantee timely advice on migration, emulation or preservation activities.

In terms of answering the Project specifics set out by UCL for this Project, LIFE feels that it has answered the following;

- What are the long-term costs of preserving digital material?

All exemplars have provided costs, where possible, for long-term preservation. VDEP, Web Archiving and UCL e-journals have all delivered lifecycle figures.

Where LIFE was not able to provide specific real costs, informed judgements have been used to complete lifecycles. (See Case Studies)

- Who is going to do it?

The LIFE Project leaves this question open for further work. The VDEP and Web Archiving Case Studies point to developing national standards and collaboration on development as the most cost-effective approach. However, from the UCL e-journals exemplar, the answer is not so clear due to the specific HE/FE research roles and responsibilities. This difference between a national library approach of custodianship and an HE/FE approach to research and access are two areas which require further discussion (see UCL e-journal conclusions, section 7.4).

- What are the long-term costs for a library in HE/FE to partner with another institution to carry out long-term archiving?

The UCL e-journals study has shown that long-term archiving within HE/FE libraries is a complicated area. Although UCL is well placed to develop a strategy if needed, major investment and changes would be required to do so. Further work is required here to explore this further.
(See UCL e-journals conclusions, section 7.4)

- What are the comparative long-term costs of a paper and digital copy of the same publication?

LIFE was unable to compare paper vs. digital long-term costs. The many different environmental, economic and geographic aspects to this question must be considered alongside the preservation costs discovered by LIFE. This is a key area for the future development of the Lifecycle Model. (See Future work, chapter 11)

- At what point will there be sufficient confidence in the stability and maturity of digital preservation to switch from paper to digital for publications available in parallel formats?

The finding that no obsolete file formats were discovered in three diverse collections has led LIFE to believe that confidence is rising in this area. LIFE did not reach the conclusion that the decision to acquire or select content based on paper or digital would be feasible, but does feel that it is now the time to have the debates within institutions. There are many benefits to switching to digital delivery from an acquisition and access viewpoint, but there are still many concerns around storage and preservation. Now that LIFE has delivered the model to use, more real data needs to be gathered to establish a clearer comparison. (See Future work, chapter 11)

- What are the relative risks of digital versus paper archiving?

The major risks identified in the LIFE Case Studies are a lack of real data, the cost of implementing a new system and workflow, the lack of long-term preservation strategies and minimal tool development as the main areas of concern when compared to paper archiving. (See future work, chapter 11)

11. Future work- LIFE2

The LIFE Project has identified the following five key themes that are strongly encouraged to be taken up in future studies in this area.

Refining the Lifecycle Model

Applying the model to more real life collections will allow further testing and refinement, and will provide the opportunity to draw more detailed conclusions from the resulting analysis. With more data, the possibility of developing predictive models for costing each of the elements can be explored. Further work on environmental, support and management costs will also refine the model and provide a more detailed context to direct lifecycle costing.

Refining the Generic LIFE Preservation Model

The work created for LIFE in the area of preservation holds much hope for future research. The opportunity exists to collect more real life preservation data and utilise this in the refinement of the model, alongside the possibility of enhancing a range of specific aspects of the model as described in section 8.8.

A National tiered repository and HE/FE

Further work is required to ascertain the areas of common interest and/or specific independence and how a national tiered repository might work.

Paper vs. Digital selection

LIFE recommends further work in this area. Now that some costs have been put in place by the Lifecycle Model, comparison between analogue and digital lifecycles can be taken further to aid institutional decisions around selection and acquisition.

Institutional file format review/technology watch

The finding that no obsolete file formats were discovered within LIFE needs to be explored. Risk analysis of archives and up-to-date technology watch information are recommended to try to build a picture of the risk level and timeframes to develop digital preservation standards.

12. Acknowledgements

The LIFE Project would like to extend its thanks and appreciation to everyone who has helped this Project reach its aims.

In particular LIFE would like to thank people in the following project areas

The Literature review

Thanks to James Watson for his comprehensive review.

The AqIMAcS model

Thanks to Andy Stephens and Helen Shenton for their work developing the model.

The Generic LIFE Preservation Model (P)

Paul Wheatley would like to thank Angela Dappert for her invaluable contributions to the development of the model, as well as Adam Farquhar and the BL eIS Architecture team for their comments and review. Thanks also to Andrea Goethals, Harvard University (formerly FCLA).

VDEP Case Study

Rory McLeod would like to thank Andy Davis, Caroline Brazier, Richard Masters and Sharon Johnson from The British Library for their help compiling this case study.

Web Archiving Case Study

Paul Wheatley would like to thank Mark Middleton, Philip Beresford, Alison Hill, Arun Persad, Ravish Mistry and Nicola Johnson.

Not only did they make this Case Study possible, they were more than happy to open up their everyday work to the outside world. Given that their operations are at a very early stage and will develop considerably over the coming years, this should be applauded. Thanks also to Arun Persad for providing the Web Archiving workflow diagram.

UCL e-journals Case Study

Paul Ayris would like thank the VDEP (Voluntary Deposit) acquisitions team, the e-media cataloguing team and the DOM (Digital Object Management) system team in the British Library for their advice and support in pursuing the Case Studies. UCL must also thank Oded Scharfstein, then Product Manager for DigiTool (the Ex Libris Digital Asset Management software) and now Vice-President in Ex Libris for Asia-Pacific. As Director of UCL Library Services, I have worked now with Ex Libris and Oded Scharfstein on two major projects. It has always been a privilege and pleasure to work with so energetic a company. In UCL, many colleagues have helped in the

Case Studies, most notably Karen Jeger, Anna Sansome, Lesley Rogers and Martin Moyle, Science Team Leader. UCL Library Services wishes to thank Steven Hall, Journal Sales and Marketing Director, Blackwells Publishing Ltd, and his colleagues for their support of the LIFE Project in terms of supplying exemplar data for Blackwells journals; and also Mark Patterson, Director of Publishing, Public Library of Science, and his colleagues for working with UCL on this Project.

The LIFE Project would also like to thank Neil Beagrie and Helen Hockx-Yu for their continued support and advice in all aspects of the Project.

Thanks also to both Henry Girling at the BL and to Christopher Carrington at UCL for their help in organising the conference and the web-site.

13. Glossary of abbreviations

A&I	-	Abstracting and Indexing
Ac	-	Access
AIP	-	Archival Information Package
Aq	-	Acquisition
BCT	-	Base Cost of Testing a Preservation Action per Object
BL	-	The British Library
BLE	-	Base Life Expectancy
BL eIS	-	British Library e Information Systems
BMP	-	Bit-Mapped Graphics Format
CAMiLEON	-	Creative Archiving at Michigan and Leeds Emulating the Old on the New
CEDARS	-	CURL Exemplars in Digital Archives
COA	-	Cost of Available Tool
CRS	-	Cost of a New Rendering Solution
DOM	-	Digital Object Management system at the BL
EISSN	-	Electronic International Standard Serial Number
ETA	-	Ending Proportion of Tool Availability
EU	-	European Union
FCLA	-	Florida Centre for Library Automation
FCX	-	File Format Complexity
FE	-	Further Education (UK)
FP	-	EU Framework Programmes
gb	-	Gigabyte(s)
GIF	-	Graphics Interchange Format
HE	-	Higher Education (UK)
HEFCE	-	Higher Education Funding Council for England
HERA	-	Higher Education Role Analysis
HR	-	Human Resources
HTML	-	Hypertext Markup Language
HVM	-	High Volume Migration Cost per Object
I	-	Ingest
ILL	-	Inter-Library Loan(s)
ILS	-	Integrated Library System
IP	-	Internet Protocol
IPR	-	Intellectual Property Rights
ISSN	-	International Standard Serial Number
IT	-	Information Technology
JHOVE	-	JSTOR/Harvard Object Validation Environment
JISC	-	Joint Information Systems Committee
kb	-	Kilobyte(s)
KB	-	Koninklijke Bibliotheek
L	-	Lifecycle costs
LCC	-	Lifcycle costing
LIFE	-	Lifcycle Information For E-Literature

LOCKSS	-	Lots of Copies Keeps Stuff Safe
M	-	Metadata
mb	-	Megabyte(s)
METS	-	Metadata Encoding and Transmission Standard
MIME	-	Multipurpose Internet Mail Extensions
N	-	Number
N/A	-	Not applicable
NESLI	-	National Electronic Site Licensing Initiative
OAIS	-	Open Archival Information System
OPAC	-	Online Public Access Catalogue
P	-	Preservation
PCP	-	Per Object Cost of Preservation
PCX	-	a graphics file Format for PCs
PDF	-	Portable Document Format
PDI	-	Preservation Description Information
PLoS	-	Public Library of Science
PNG	-	Portable Network Graphic
PON	-	Proportion of Normalisation
PPA	-	Performing Preservation Action
PREMIS	-	Preservation Metadata Implementation Strategies
PTA	-	Proportion of Tool Availability
QA	-	Quality Assurance
QAA	-	Quality Assurance Actions
RAE	-	Research Assessment Exercise
S	-	Storage
SCM	-	Setup Cost of Migration
SCONUL	-	Society of College, National and University Libraries
STA	-	Starting Proportion of Tool Availability
T	-	Time
TB	-	Terabyte(s)
TDC	-	Tool Development Cost
TEW	-	Technology Watch per format
TIFF	-	Tagged Image File Format
TLSS	-	Teaching and Learning Support Section, UCL Library Services
txt	-	ASCII text files
UCL	-	UCL (University College London)
UKWAC	-	UK Web Archiving Consortium
ULE	-	Unaided Life Expectancy
UME	-	Update Metadata
URL	-	Uniform Resource Locator
VAT	-	Value Added Tax
VDEP	-	Voluntary Deposit collections at the British Library
VLE	-	Virtual Learning Environment
VS	-	Versus
WMF	-	Windows Metafile Format

XML - Extensible Markup Language