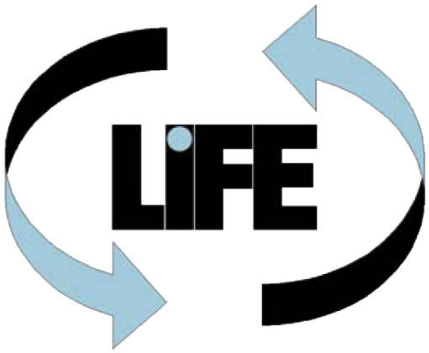


The LIFE^2 Final Project Report

A JISC-funded joint venture project under 03/06, Repositories and Preservation Capital Programme, and supported by the LIBER Access and Preservation Divisions

A NOTE ON THE AUTHORS

This report had contributions from a range of people, but was primary written and reviewed by the LIFE Project Team:

- Paul Ayris, Director of UCL Library Services (UCL)
- Richard Davies, LIFE^2 Project Manager (BL)
- Rory McLeod, Digital Preservation Manager (BL)
- Rui Miao, LIFE^2 Project Assistant (BL)
- Helen Shenton, Head of Collection Care (BL)
- Paul Wheatley, Digital Preservation Manager (BL)

Acknowledgement and thanks go to the following people who wrote particular sections of the report:

- Section 3: The Models – Paul Wheatley (BL)
- Section 4: SHERPA DP Case Study – Stephen Grace (CeRch) and Gareth Knight (CeRch)
- Section 4: SHERPA-LEAP Case Study – Rebecca Stockley (SHERPA-LEAP), Martin Moyle (UCL), Jacqueline Cooke (Goldsmiths), and Adrian Machiraju (Royal Holloway).
- Section 5: British Library Newspapers Case Study – Rory McLeod (BL)

The report was edited by Richard Davies (BL).

Final Report (22/08/08)

OVERVIEW OF CONTENTS

CONTENTS PAGE

LIST OF TABLES

LIST OF FIGURES

1 EXECUTIVE SUMMARY

The first phase of LIFE (Lifecycle Information For E-Literature) made a major contribution to understanding the long-term costs of digital preservation; an essential step in helping institutions plan for the future. The LIFE work models the digital lifecycle and calculates the costs of preserving digital information for future years. Organisations can apply this process in order to understand costs and plan effectively for the preservation of their digital collections

The second phase of the LIFE Project, LIFE[2], has refined the LIFE Model adding three new exemplar Case Studies to further build upon LIFE[1]. LIFE[2] is an 18-month JISC-funded project between UCL (University College London) and The British Library (BL), supported by the LIBER Access and Preservation Divisions. LIFE[2] began in March 2007, and completed in August 2008.

The LIFE approach has been validated by a full independent economic review and has successfully produced an updated lifecycle costing model (LIFE Model v2) and digital preservation costing model (GPM v1.1). The LIFE Model has been tested with three further Case Studies including institutional repositories (SHERPA-LEAP), digital preservation services (SHERPA DP) and a comparison of analogue and digital collections (British Library Newspapers). These Case Studies were useful for scenario building and have fed back into both the LIFE Model and the LIFE Methodology.

The experiences of implementing the Case Studies indicated that enhancements made to the LIFE Methodology, Model and associated tools have simplified the costing process. Mapping a specific lifecycle to the LIFE Model isn't always a straightforward process. The revised and more detailed Model has reduced ambiguity. The costing templates, which were refined throughout the process of developing the Case Studies, ensure clear articulation of both working and cost figures, and facilitate comparative analysis between different lifecycles.

The LIFE work has been successfully disseminated throughout the digital preservation and HE communities. Early adopters of the work include the Royal Danish Library, State Archives and the State and University Library, Denmark as well as the LIFE[2] Project partners. Furthermore, interest in the LIFE work has not been limited to these sectors, with interest in LIFE expressed by local government, records offices, and private industry. LIFE has also provided input into the LC-JISC Blue Ribbon Task Force on the Economic Sustainability of Digital Preservation.

Moving forward our ability to cost the digital preservation lifecycle will require further investment in costing tools and models. Developments in estimative models will be needed to support planning activities, both at a collection management level and at a later preservation planning level once a collection has been acquired. In order to support these developments a greater volume of raw cost data will be required to inform and test new cost models. This volume of data cannot be supported via the Case Study approach, and the LIFE team would suggest that a software tool would provide the volume of costing data necessary to provide a truly accurate predictive model.

2 INTRODUCTION

2.1 Purpose of this Section

This section introduces the work undertaken for the second phase of the LIFE Project (LIFE[2]). It also details the origins of the project, summarising the outputs from the first phase (LIFE[1]) and linking the phases of the project moving from the LIFE[1] to LIFE[2]. The final part of this section outlines the structure of the report, the Methodology used throughout the Project and how to get the most out of the project documentation.

- **Summary of LIFE[1]** gives an overview of what the outcomes from the first LIFE Project were.
- **Uptake since LIFE[1]** outlines some of the interest there has been in adopting the LIFE Model since the completion of LIFE[1].
- **Why do we need a LIFE[2]?** expands on the reasoning behind this second phase of the project.
- **Project Aims** lists the aims outlined in the project plan for LIFE[2] and what has been done to fulfil those aims.
- **Structure of Report** explains the framework of the report and how the sections link together.
- **LIFE Methodology** gives an outline of the methodology used throughout the project.
- **How to use LIFE** briefly outlines how to get the most out of this report from a number of different perspectives

However, if a summary of LIFE[2] is needed, a better starting point would be the LIFE[2] Project Summary produced, which is available online (www.life.ac.uk).

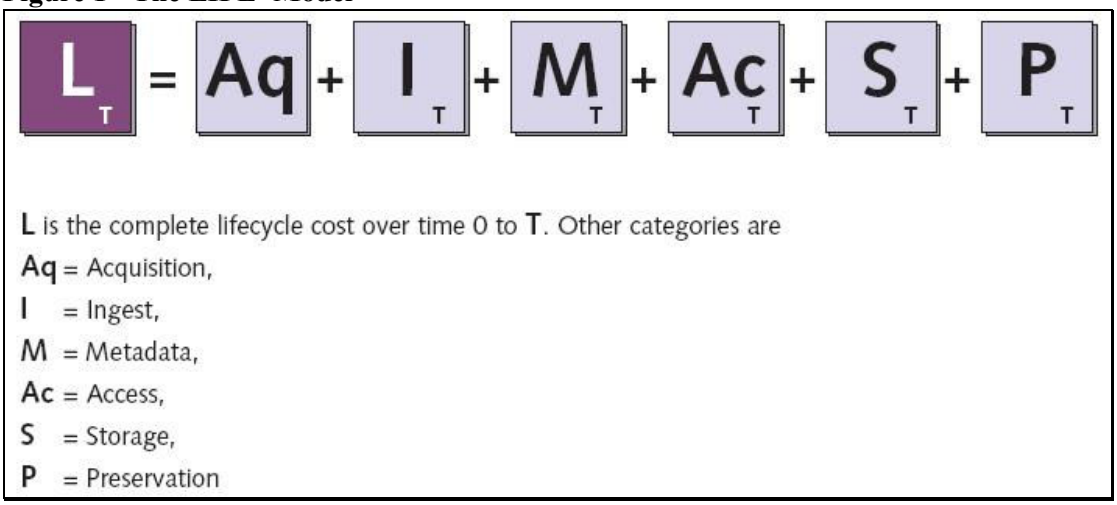
2.2 Summary of LIFE[1]

The first phase of the LIFE Project ("LIFE[1]") drew to a close with an international conference in April 2006. A full description of the project and all supporting project documentation is still available online (www.life.ac.uk/1).

Run from 2005 to 2006, the first phase of LIFE (Lifecycle Information for E-Literature) project, LIFE[1], has made a major contribution to understanding the long-term costs of digital preservation, an essential step in helping institutions plan for the future. Based on a comprehensive review of existing lifecycle models and digital preservations, LIFE[1] project has developed a lifecycle based methodology to calculate the costs of preserving digital information for the next 5, 10 or 100 years.

The LIFE Model v1 broke down a digital object's lifecycle into six main lifecycle categories as summarised in
Figure 1. Calculating the summation of these elements over a specific time period provides a complete lifecycle cost.

Figure 1 - The LIFE[1] Model



A further breakdown of the lifecycle categories and elements, as well as a full explanation of each element is provided in the LIFE[1] Project Final Report (Section 4)

2.2.1 Generic Preservation Model v1

Due to the lack of work done in the areas of digital preservation costing before 2005, LIFE[1] also produced the Generic Preservation Model to further develop the Preservation element of the model, using which institutions can start to reduce the spikes of cost, as well as the frequency of the preservation actions.

In the Generic Preservation Model, key elements of preservation activities were identified and the factors which contributed to their costs were modelled. These included elements such as the Proportion of Tool Availability, Tool Development Costs and Format Complexity. A spreadsheet tool for calculating the costs for digital objects of varying file formats was also developed as part of the model. A detailed introduction on the Generic Preservation Model can be found from the LIFE[1] project final report (Section 8), and the spreadsheet is available for downloading from the LIFE website.

2.2.2 Case Studies and Findings from LIFE[1]

To test and evaluate the LIFE methodology, LIFE[1] project chose three case studies: Web Archiving, Voluntarily-Deposited Electronic Publications (VDEP) at the British Library, and E-Journals at UCL. By using these Case Studies, which were vastly different in both content and workflow, key costs were identified for each element in the lifecycle, enabling the estimation of costs for a single title, item or instance over a given time period.

2.2.2.1 Web Archiving

The Web Archiving Case Study considered the costs of the British Library's web archiving activities, which selected and archived around 1000 website instances each year. In this Case Study, the cost of preserving web materials was found to be high, particularly in the short term. Preservation represents approximately 55% of the complete lifecycle costs. 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.

Greater efficiencies, and the introduction of more automated processes, will reduce Web Archiving costs considerably, but manual effort is likely to leave the costs of Ingest at a relatively high level for the medium term. The Case Study suggested the introduction of legal deposit legislation covering web materials would dramatically cut the cost of the IPR (Intellectual Property rights) portion of the acquisition costs. The full report of Web Archiving Case Study can be found from the LIFE[1] Project Final Report (Section 6).

2.2.2.2 E-Journals

The E-journals Case Study was based at UCL Library Services. At the time of the Case Study, 8668 e-journal titles were logged in a UCL Access database. With the emphasis on giving access to e-journal literature, instead of Ingest, Storage or Preservation, the e-journal Case Study found that different elements of the Lifecycle Model fell under the spotlight when UCL analysed its own workflows and processes. It was also noted that for most HE libraries, activity-based costing is not yet embedded in the workflow of the organisation.

In terms of the lifecycle, 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. One aspect of the E-journals Case Study, however, is significant for HE institutions to consider whether university libraries each should be responsible for digitally preserving their own e-journals or it is more cost effective to have this activity performed by a trusted third party. The full report of E-Journal Case Study can be found from the LIFE[1] Project Final Report (Section 7)

2.2.2.3 VDEP

Voluntarily-Deposited Electronic Publications (VDEP) housed at the BL provided the final Case Study and involved the analysis of over 230,000 files. Using the LIFE Model, VDEP preservation costs are projected to go down over time. At the time of the Case Study, there were no obsolete file formats within VDEP and indeed LIFE[1] struggled to find any formats at risk in any of its three Case Studies.

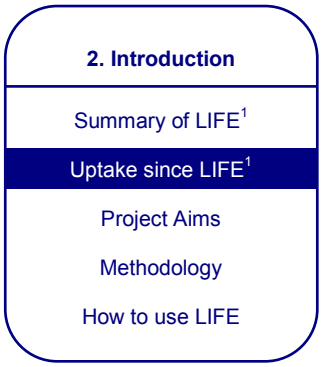
In VDEP Case Study, both Ingest and Metadata processes are currently very manual and in their present form incur a high proportion of the lifecycle cost. The LIFE[1] study showed that large-scale investment at the Ingest point to automate metadata would vastly reduce processing costs. The full report of VDEP Case Study report can be found from the LIFE[1] project final report (Section 5).

2.2.2.4 LIFE[1] Overall

The three 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. More detailed practical and strategic findings for each of the Case Studies can be found in the LIFE[1] project final report (Section 9)

Meanwhile, since LIFE[1] focused on the development of a model to estimate the long-term preservation costs, the Case Studies considered by the project did not contain activities addressing the preservation of content, such as technology watch, preservation planning or migration. These aspects have been further addressed and refined in the updated LIFE Model in LIFE[2].

2.3 Uptake since LIFE[1]



Since the LIFE[1] Conference (April 2006) and throughout LIFE[2], there has been considerable interest in the LIFE work, outside the original project partners. The interest in adopting the LIFE work gave the team invaluable feedback when it came to updating the model to its current version.

The uptake of the LIFE work and the feedback gained from this process is an essential part to LIFE[2]. Testing the model with these additional Case Studies adds to the robustness of the LIFE work. Part of the documentation for this stage of the project, aims to make adopting the model as easy a process as possible. This should be evident throughout this report in a number of areas. The lifecycle stage definitions have been fully reviewed, with examples to make the process easier to understand. The exemplar Case Studies have also been written so that other institutions (whether libraries, museums, archives or other interested parties) can easily extract the practical aspects of adopting the LIFE model.

The LIFE work has been successfully disseminated throughout the digital preservation and HE communities. Early adopters of the work include the Royal Danish Library, State Archives and the State and University Library, Denmark as well as the LIFE[2] Project partners. Furthermore, interest in the LIFE work has not been limited to these sectors, with interest in LIFE expressed by local government, records offices, and private industry.

2.4 Why do we need a LIFE[2]?

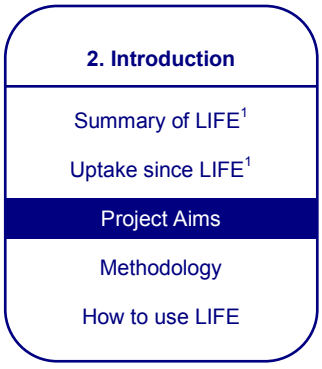
While both the conference and project as a whole were seen to be very successful, the project team felt that there was clearly further work that needed to be done in this area. This need for further development was also mirrored by comments from conference delegates.

The LIFE team successfully applied for a second phase of the project (LIFE[2]) which resulted in the work documented in this project report. Essentially, this second phase revolved around a few key themes:

- a thorough testing of the economic validity of the LIFE model;
- further development of the model with a wider range of Case Studies (including non born-digital material);
- an assessment of the analogue versus digital lifecycle costs, mapping both to the LIFE methodology.

These key themes were developed further into the Aims of the project.

2.5 Project Aims



These are the stated aims of the project as outlined in the initial project funding bid and project plan. Progress made is highlighted under each aim.

1. Refine the LIFE methodology for the analysis and costing of the lifecycle of digital objects;
   - The LIFE Model has been refined into version 2 of the model (Section 3.4)

2. Provide a cross section of exemplar Case Studies, both to inform the LIFE methodology and to provide a benchmark for comparison and evaluation;
   - ► The three Case Studies are discussed in Sections 4 and 5.

3. Enable HE and FE institutions to apply the LIFE methodology simply and easily to their own collections, and thus to evaluate and compare their activities in order to inform planning and increase workflow efficiency;
   - ► The LIFE team have worked within the HE community to disseminate the model and to ease adoption of the model. Workflows have also been developed for the Case Studies to ease adoption.

4. Compare, contrast and analyse the lifecycle costs of paper and digital collections, informing the use of differing approaches to preservation and access via digital and other surrogate technologies;
   - ► The British Library Newspapers Case Study (Section 5) successfully compares the lifecycle costs of both analogue and digital collections.

5. Identify where efficiencies can be made in the lifecycle costs of digital materials and provide guidance to funding bodies in areas such as preservation services and preservation tools;
   - ► All the three Case Studies (in particular the SHERPA DP Case Study in Section 4) examine the efficiencies of a variety of digital collections across the lifecycle.

6. Disseminate project findings and enable take up of the LIFE methodology.
   - ► This report, along with all project outputs will be made publicly available through the LIFE website www.life.ac.uk.

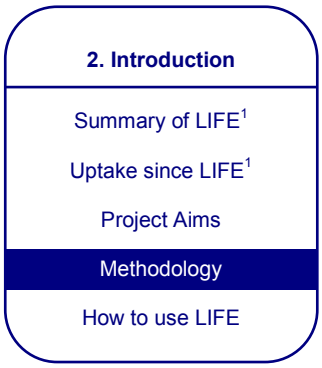
2.6 Structure of Report

As outlined in the Contents Page the report follows the following structure:

- ► **Section 3** details the LIFE Models and the economic background to the project.
- ► **Section 4** analyses the Institutional Case Studies (SHERPA DP and LEAP)
- ► **Section 5** analyses The British Library Newspapers Case Study which compares analogue and digital lifecycles
- ► **Section 6** brings the Case Study findings and conclusions from each of the Project outputs
- ► **Section 7** discusses possible areas of future development
- ► **Section 8** lists all the acronyms used throughout the report.

Each section starts with a brief outline how it might be useful. Each section can be read independently of each other. However, to gain a complete understanding of the project and its results it would be advisable to at least check through the entire report. This is particularly relevant to the sections on the Case Studies. For example, by simply reading conclusions section of the report and not the Case Study write-ups themselves, a great deal of the context will be missed and the resulting conclusions will be less meaningful.

2.7 Methodology



As with LIFE[1], a Case Study approach was chosen for this second phase of the LIFE Project. The Case Studies discussed in Sections 4 and 5 were chosen in order to both maximise the feedback gained and to test the model thoroughly. However, throughout these different Case Studies it was important to have a consistent approach to the work done.

This section outlines the LIFE Methodology, both to demonstrate how the work was completed, but also as an indication of how to adopt the LIFE Model for those institutions wishing to do so. In this way, this section also ties in with the following section ('How to use LIFE').

Once the independent economic review was completed, the LIFE Model used in the first phase of the project was thoroughly reviewed and updated. As was outlined in the Model documentation earlier in this section, feedback was received from a number of sources (including the HE and digital preservation communities) all of which fed into a working version of the Model – v1.1.

The document outlining v1.1 of the model was published for further comment, and then used as the working model for the Case Studies. Each of the Case Study partners used this document as starting point for the new Case Studies. To guide each of the Case Studies, two templates were also developed.

The following sections expand on the methodology that the LIFE team adopted:

- Outline of the LIFE Methodology
- The Case Study document template
- Capturing Costs
- The LIFE Costing spreadsheet template

2.7.1.1 Outline of Methodology

As discussed in the Case Study feedback, the LIFE methodology adapted and changed as feedback was received. While the team felt that it was important to have a structure in place for the Case Studies to allow for a consistent approach across the LIFE work, there is nonetheless flexibility within the methodology to allow for differing collections as well as institutional differences.

Key parts of the methodology are elaborated on in the next sections, however broadly the methodology follows these stages:

1 Identity collection and timeframe for Case Study.

2 Review LIFE Model v2
The section on the LIFE Model (page17) can be viewed separately from the rest of the report, and therefore should be used as a starting point for all queries relating to the LIFE Model and the breakdown of costs and processes.

3 Identify key staff involved in each Stage and Element of the lifecycle.
This might include staff working with the collections, as well as institutional repository staff, administrators, finance, estates, and preservation staff.

4 Interview staff identified about how the LIFE terminology fits with existing terms and process. This will help to identify any potential problematic areas (for examples conflicts with existing process or terminology with LIFE).

5 Feedback on Issues
An important role of the Case Studies is to feed any issues or conflicts with the Case Studies into the Model and an updated methodology.

6 Development of Workflow
Production of workflow diagrams is a useful process for as both an overview of the process, as well as for identifying potential issues.

7 Cost analysis to LIFE Model
This involves capturing and categorising the entire lifecycle for Year 1. With the current Case Studies this was largely based on activity-based costing.

8 Populating LIFE Model spreadsheet
A spreadsheet with the LIFE Model mapped to it can be used for Year 1 through to Year 10 of a collection. However, making accurate predictions for the lifecycle is still very much in its infancy.

9 Pulling it together
Once costings have been inputted into the spreadsheets, graphical summaries can be produced to identify spikes in activity and cost. Combining this analysis with the mapping of the processes completed earlier will help identity an overall picture of the collections lifecycle.

The two tables below outline some examples of the costing terms used in the Methodology and throughout the model.

Table of Costing Terms with Examples

Type of Cost	Explanation	Example in Model
Staff (or Labour)	An activity that requires time and effort from a member of staff. For the LIFE^2 Case Studies this was often measured by activity-based costing	The creation of a Submission Agreement (within the Acquisition Stage) requires a person to write that agreement
Hardware	Cost of purchasing IT hardware. This could be a one-off or ongoing cost. Would be separate from any costs associated with supporting the hardware.	Cost of storage hardware within the Storage Provision Element (Bit-Stream Preservation Stage)
Software	Cost of IT software. This could be a one-off or ongoing cost if software updates were required	Cost of Repository Software.
Capital / Estates	These might be one-off or recurring costs to do with land, infrastructure or building costs	Can be included within Non-lifecycle Costs section or included in FEC salary costs (i.e. spread across the lifecycle).
Lifecycle Cost	A cost that is reflected in the lifecycle stages and elements of the LIFE Model. These might be one-off or recurring costs.	Creation or Purchase costs
Non-lifecycle Cost	A cost that is considered outside of an object's lifecycle. However, what is termed 'lifecycle' and 'non-lifecycle' cost will vary between institutions. For example some institutions may wish to include Repository Administration within the lifecycle, some may not.	Management and Administrative costs for an institution. This might also include building (estates) costs, and systems infrastructure.

Activity-based Costing Explanation

Method of Costing	Explanation	Category Example
Activity-based Costing	Costing based on the amount of time spent on a particular activity (in the case of LIFE – on a sub-element level) and the full salary cost of that individual. It can be calculated as a percentage of an individual (or team's) time (as with the Newspapers Case Study) or as a fixed amount of time (e.g. 3 days a week, as with the SHERPA DP Case Study)	A great deal of the Case Study staff costs were based on Activity-based costing. All the spreadsheets give details of how each calculation based on activity-based costing was arrived at.

2.7.2 Case Study Template

A template structure was produced to guide Case Study partners in their write-ups. This was a flexible and fairly simple Microsoft Word document that outlined the need for certain key pieces of information:

- Background to the Institution and the Case Study
- A walk through of the LIFE Model by Stage, Element and Sub-Element designed to capture not only the costs, but also the specific process for each institution, and any differences in terminology
- Feedback on the LIFE Model to support into the final review (which would then become v2)
- Cost Results and Conclusions
- Any comments on the process of adopting the LIFE Model

This structure and the areas covered evolved over the period of the Case Studies, but it allowed the LIFE team to ensure that there was a consistent approach in terms of the information that they were receiving and the questions that were being asked. It also proved to be a helpful starting point for the Case Studies.

2.7.3 Capturing the Costs

LIFE implemented a simple methodology for the capture, calculation and recording of lifecycle costs. Key costs were identified for each element in the lifecycle. These might include equipment costs, setup costs and ongoing staff costs. An appropriate method of capturing these key costs was then identified and applied. Capital costs were averaged across their expected lifetime and the numbers of objects that would be processed. Staff costs were captured using studies of the involved personnel and the time they spent on different tasks (activity-based costing). Costs were simply projected over time based on present day value, without consideration for inflation. LIFE calculated costs for 1, 5, and 10 years.

As outlined later on, in the section defining the LIFE Model (page 17), costs can be incurred at each stage of the lifecycle. Costs may be incurred just once, may accrue over time, or recur on a regular or irregular basis. The Case Studies highlight this cross section of cost types including one-off costs in the first year for content Selection, costs that accrue over time such as Storage Provision, and recurring costs for Preservation. The methodology enables the estimation of costs for a single title, item or instance over a given time period.

2.7.4 Costing Spreadsheets

The second template used was a Microsoft Excel Costing Spreadsheet populated with the LIFE Model v1.1. Again, this template evolved slightly over the process of completing the Case Studies, but it allowed the LIFE team to have a consistent approach to what information they were capturing, and how to go about the costing exercise.

It is important to note that the spreadsheet template was not simply used to capture the lifecycle costs from each Case Study. It was critical to the success of the project that the team not only collected the lifecycle costs but also the processes and activities behind those costs. This allowed the team to make the data meaningful.

Each of the spreadsheets now contain certain key pieces of information:

Lifecycle Stages Sheet

This sheet gives all the costings on a Stage, Element, and Sub-element level. It includes the way the costing was calculated, as well as a practical explanation of the process for each sub-element for each particular collection. For those people interested in detailed costs, this sheet will be of most use. For those more interested in the overall costs, the summary sheet will probably be of more use. By hovering the mouse over the LIFE terms, the full definition will appear in a pop-up box.

Acronym & Staff Cost Sheet

This sheet contained details of the higher-level costs used in the calculations. This would include, for example, the overall staff salary levels for various positions. These costs are then linked to the calculations on the remaining sheets. So, for example, a user can edit the overall staff costs in this sheet, and the lifecycle costs on the other sheets will be updated automatically.

Summary Sheet

This sheet gives an overall snapshot of the costs, as well as a summary graph of the lifecycle. As with the other calculations, it is also linked to the other sheets, so any data will automatically update if any changes are made to the calculations on other sheets.

Each of the spreadsheets for the Case Studies is available from the LIFE website and can be edited and adapted to examine other institutional costs. A great deal of the costs involved in the lifecycle of the collections are based on activity-based costing, which is worth looking at in a little more detail.

The linking between the costs and the spreadsheet calculations is also worth explaining a little more. Figure 2 gives a snapshot of one of the spreadsheets from the Newspapers Case Study.

Column A gives the Stages, Elements and Sub-elements from the LIFE Model. When the mouse is moved over each of the terms, a pop-up box appears with a definition of that term as outlined in the LIFE Model (Section 3.4, page 17).

Column B gives a practical explanation of what that actually means in practice for each individual Case Study.

Column C details what the cost calculation is to arrive at the overall costs for each activity given in Columns F to Q (which covers Years 1 to 10).

Column D gives a percentage of time that this activity is deemed to take if it is based on what would be termed ‘activity-based costing’. This additional information in columns C and D is

useful for the activity-based costing on which a great deal of the lifecycle costs are based. The template is such that an institution can alter the labour costs that are specified in a salary table on a separate sheet and these updated costs will filter through to all the costing calculations.

Figure 2 - Snapshot of Burney Digital Spreadsheet

	A	B	C	D	E	F	G	H
1								
2						**Burney Digital Collection- Lifecycle Pr**		
3						**v 2.0**		
4								
5	**Lifecycle Stages**	**Practical Explanation**	**Cost notes**	**% of Staff time**	**Year 1 (individual costs)**	**Year 1**	**Year 2**	**Year 3**
6								
7	**Creation or Purchase**							
8						0.00	0.00	0.00
9		Cost to digitise the total archive plus create associated project information	creation of digital archive			189,710.00	129,373.00	129,373.00
10								
11	**Sub Total**					189,710.00	129,373.00	129,373.00
12								
13	**Acquisition**							
14	**Selection**							
15	Selection Policy (policy/procedure)		10% Curatorial Grade B	10.0%	5,556.32	5,556.32	0.00	0.00
16	Selection (action)	Sort Microfilm into correct order	50% Curatorial Grade C, 6 months	25.0%	10,713.15	10,713.15	0.00	0.00
17	Selection Metadata (metadata)							
18	**Submission Agreement**							
19	Submission Agreement (policy/procedure)	Setting up of contract with third pa	10% project Grade A	10.0%	6,647.62	6,647.62	6,647.62	6,647.62
20	Negotiation of Submission (action)	Communication and negotiation with producers/depositors.	5% Legal support Grade A	5.0%	3,323.81	3,323.81	0.00	0.00
21	Submission Metadata (documentation)							
22	**IPR & Licensing**							
23	IPR & Licensing (policy/procedure)	Management of contract agreement with third party	5% Product development Grade B	5.0%	2,778.16	2,778.16	0.00	0.00
24	Negotiation of Rights (action)		5% Product development Grade B	5.0%	2,778.16	2,778.16	0.00	0.00

For example, Row 16 gives the costs for 'Selection (action)'. The team knows from the cell F16 that the cost in Year 1 is £10,713, but how is this cost calculated? Cell C16 shows that this activity took 50% of a Curatorial member of staff at Grade C for 6 months. This would mean 25% of that person's time over a one-year period, which is what is inserted into the '% of Staff Time' column in cell D16. Thus, the final cost in F16 is not simply a number, it is a calculation of that 25% multiplied by the annual salary cost for a Grade C that is specified on another sheet.

Having these costs calculated automatically, allows for the updating of cost information with relative ease. It also allows for a greater understanding of how these costs were calculated.

For example, supposing the salary scales were changed, and an institution wanted to see how that would change the lifecycle costs for a collection. Rather than having to go into the spreadsheet and update every cost individually, all that is needed is to update the master salary table and the updated lifecycle costs will be automatically calculated.

This spreadsheet template aimed to remove some of the work for Case Study authors, as well as remove the complexity of the spreadsheets as far as possible. Once they were completed, we wanted to be able to have all the costings included in the spreadsheets, so that they were available is needed, while at the same time having the summary sheet there to give more of an overview of the lifecycle.

2.7.5 Case Study Feedback on the Methodology

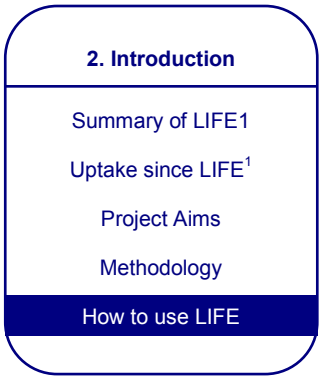
As with the first phase of LIFE, the process of setting up and carrying out the three diverse LIFE² Case Studies fed back into the Methodology used. With both the Newspapers and

Repository Case Studies it was essential the team had a consistent approach. However, as each of the Case Studies progressed, so the method employed had to be adapted and changed. This was particularly the case with the Newspapers Case Study when analysing the usefulness of the LIFE Model for analogue collections.

For the Newspapers Case Study, the level of analysis that was required to identify the costs associated with an analogue collection was not insignificant. Considerable effort was required to produce detailed business analysis of the functions and costs that a large analogue collection entails. This analysis provided a strong challenge to the methodology which had until this time been used solely for digital collections. This has led to a much tighter definition of the methodology and the steps that are mandatory to produce consistent results.

There were also different challenges when dealing with external institutions for the Case Studies. Processes such as identifying the correct people to deal with and gaining an understanding of another institution's processes require considerable time and resources. Indeed, far more effort was required to implement the LIFE^2 Case Studies than was expected, and this placed a considerable strain on project resources. There is also a level of sensitivity that needs to be observed when examining costs of this nature. As discussed further in the concluding comments, there is a fine line between costing analysis and audit.

2.8 How to Use LIFE



One of the key considerations for LIFE^2 is to make the LIFE model and findings more accessible to those institutions wishing to either adopt the model, or to make use of the findings - essentially, to try to answer the question, 'How is the LIFE work useful for our own collections?'

With this in mind, this brief section outlines the sections that would be particularly useful for institutions with an existing repository which wishes to add a new content stream, or possibly for an institution considering setting-up a new repository.

The three SHERPA-LEAP repository Case Studies (Section 4.4) will be of particular interest here. UCL (page 68), Royal Holloway, University of London (page 63) and Goldsmiths, University of London (page 55) all have existing ePrints digital repositories. Each repository Case Study gives details of what the costs were across the entire lifecycle of particular collections. This includes information on not only what the costs for these institutions were, but how they went about measuring these costs through activities such as activity-based costing.

Ultimately, the team's aim with LIFE is for the discussion in these Case Studies to help answer the following questions:

- Where are the spikes in cost when adding a new content stream?
- How can we reduce these activities and costs?
- How can we predict what the future cost of adding a new collection might be?

For those institutions wishing to compare costs across analogue and digital collections, the Newspapers Case Study (Section 1, page 75) gives a guide to mapping out these costs and where the different cost spikes may occur.

Before going through any of the Case Studies however, a familiarity with the LIFE Model itself is needed. Sections 3.4 on the LIFE Model and 3.5 on the Generic Preservation Model

can be read independently of the rest of the project report, and should therefore be the starting point before examining the Case Studies.

In terms of an institution with no existing digital repository seeking guidance on the costs of keeping digital collections, it is important to state that the LIFE Project does not give a step-by-step guide to setting up an institutional repository. However, it does give a guide to not only the costs involved, but also to the possible areas of activity that might be particularly resource-intensive. Again, the sections on the Models and then the Case Study write-ups are the key starting points here.

2.8.1.1 A Note on Costs and Figures

When examining the costing examples throughout the report, the following points should be observed:

- All costs are given in pounds sterling (£).
- Decimal points are represented by a full stop (.). For example £1,565,212.42 equals one million five hundred and sixty five thousand, two hundred and twelve pounds and forty two pence.
- All the costs summarised in this final report are also available in the spreadsheets available from the LIFE website. Each of the three case studies is accompanied by a separate Excel spreadsheet.

Within each of the Case Studies, the costs given have been made as a meaningful as possible. For example, not only have the lifecycle costs been given, but also the processes behind those costs, and how those costs were calculated. In each of the costing spreadsheets, an explanation of the costs is given.

For example, the cost for SHERPA DP of Ordering and Invoicing is given as £55.60. However the process this actually entails is detailed as well as how the cost is arrived at (in this case 1 hour of an Administrative Officer's time).

It is also worth noting that while the LIFE team do calculate exact costings with pounds and pence, a more meaningful way of looking at the lifecycle costs is through the graphs. As is discussed throughout the Case Studies, the final costs are accurate, but are very collection-dependent. The graphical presentations of the results allows for a more overall picture of the costs across the lifecycle of the objects analysed.

It can be misleading to take the costing in the spreadsheets as absolute. In many of the Case Studies the costings should be regarded as illustrative rather than to the penny accurate. Therefore, for your reference, the costing spreadsheets do give exact costing calculations with no alterations to the figures. However, the per-entity cost tables in this report use figures that are rounded up by at least one significant figure.

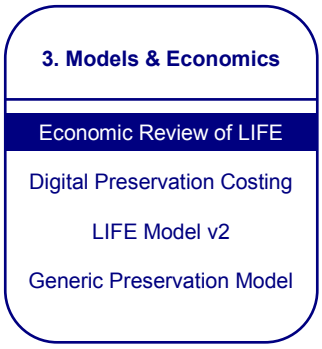
3 MODELS & ECONOMICS OF LIFE

3.1 Purpose of this Section

This section details the work undertaken for the first work package (WP1) of LIFE², including updates to the Generic Preservation Model, the LIFE Model itself and the review which Professor Bo-Christer Björk undertook on the economic aspects of LIFE.

- **An Economic Review of LIFE** outlines the report written by economist Bo-Christer Björk on the approach used for both the first and second phases of LIFE. This independent review was an essential first step in ensuring that the LIFE approach was valid.
- **Aims of Digital Preservation Costing** highlights some of the different approaches that an organisation can take to costing activities.
- **LIFE Model** describes the current version of the model (version 2) which has been thoroughly updated from the first phase of the project.
- **Generic Preservation Model** (GPM) summarizes the update to the preservation model with accompanying spreadsheet.
- **Future developments** looks at possible areas for future work for the LIFE Model

3.2 An Economic Review of LIFE



When the first phase of LIFE was completed, one of the key elements that the team wanted to work on for LIFE² was a review of the economic approach used. Professor Bo-Christer Björk from Hanken, the Swedish School of Economics and Business Administration, was brought on board to complete a full independent review to the LIFE approach.

The report largely validated the approach taken by the LIFE team. At the same time, it provided a number of recommendations to steer the second phase of the project in the right direction. The recommendations are summarised below, and the full report is available from the LIFE Website[1].

The report and models of LIFE phase 1 provide a very good starting point for the work which continues in phase 2. Professor Björk's report provides a critical reading of this work and presents some suggestions for improvements. The central ones are:

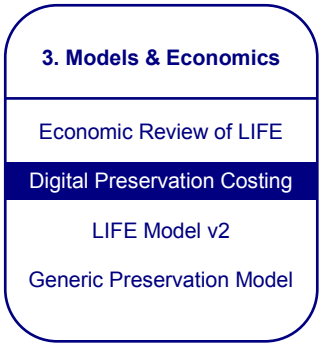
- The context for using the LIFE formulae should be further elaborated and some use cases, where the exact way in which the formulas are used to inform decisions about preservation strategy, should be developed.
- The models could be extended also to analogue material and in particular to the important issue of conversions to from analogue to digital format.

[1] Björk, B.-C. (2007) *Economic evaluation of LIFE methodology*. Research report. LIFE Project, London, UK. Available from: http://eprints.ucl.ac.uk/7684/

- The LIFE models should mainly be used for collections where the organisation assumes the primary responsibility for long-term preservation of the material.
- Long-term preservation of teaching objects should be left outside the scope of the LIFE² work.
- All calculations should be done using real-term, inflation-adjusted costs. No discounting should be applied.
- The basic unit of preservation (to be used in the cost calculation) should be clarified. For scientific journals in particular, this unit should be the article having individual metadata rather than at title level or the yearly volume.
- A taxonomy of preservation objects should be developed.
- The LIFE formulas should be further developed to take into account the comments in this report. In particular, the formulas should take into account a realistic strategy for quickly reducing the number of different file formats to a few dominant ones, as well as sharing the costs of technology watch among a larger number of players.

This validation of the LIFE approach, allowed the team to move forward with the reviews of both the LIFE Models and the further Exemplar Case Studies as planned. The review was particularly helpful in the direction of the overall LIFE Model as outlined in Section 3.4 (page 17).

3.3 Aims of Digital Preservation Costing



3.3.1 Purpose

The aim of this section is to discuss why organisations might want to cost digital preservation activities and to suggest which approaches to costing might be most useful in supporting those activities. Many of the more complex issues encountered by the LIFE team in developing lifecycle costing techniques depend to an extent upon the objective of the costing activity. It was therefore seen to be useful to identify clearly and to discuss the uses of these different costing aims or purposes.

3.3.2 Preservation Costing Aims

The aims of costing a digital preservation activity, and related analysis work, might be represented as:

1. An organisation is planning to set up a new digital repository and wants to know how much this will cost.
2. An organisation wishes to calculate the total cost of establishing its digital repository.
3. An organisation wishes to assess the cost of running its digital repository in order to compare this figure with the costs of running repositories at other organisations.
4. An organisation is considering whether to ingest a new content stream into its digital repository, and wants to know how much more this will cost on top of existing repository costs.

5. An organisation wishes to evaluate how efficiently a particular content stream is preserved in its repository
6. An organisation wishes to assess the impact in cost or efficiency of adding a new tool to its repository workflow and of changing an existing process.
7. An organisation wishes to compare the cost of analogue and digital preservation.

3.3.3 Preservation Costing Approaches

Two different approaches have been used to cost digital preservation activity:

A) Top-down audit of all preservation and repository activity[2]
B) Bottom-up lifecycle costing of activities relating to a particular content stream[3]

Table 1 - Costing Aims and Approaches

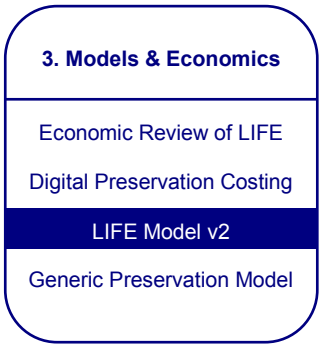
Costing Aim		Costing approach	
No.	Description	A) Top down audit	B) Bottom up lifecycle
1	Cost of new repository	Useful approach for costing this aim. Audits of existing repositories are likely to provide useful information for organisations planning for the setup of a new repository.	An inefficient way of costing for this purpose. Not very practical.
2	Complete cost of existing repository	Good approach for costing this aim.	An inefficient way of costing for this purpose. Not very practical.
3	Repository running cost	Should be possible to separate the setup and ongoing costs and produce useful results using this approach.	A costing of the lifecycle of one or more content streams would elicit at least some of the running costs and provide useful results.
4	Cost of new content stream	Difficult to assess with this approach as it would be necessary to divide up costs between different content streams that may have different preservation processes associated with them.	A useful approach. Lifecycle costs for existing repositories are likely to be useful for an organisation planning to add a new content stream. Further development of the approach to estimate the costs of a new content stream will be addressed in the proposed LIFE[3] Project.
5	Evaluate content stream efficiency	Difficult to assess with this approach as it would be necessary to divide up costs between different content streams that may have different preservation processes associated with them.	The Lifecycle approach is appropriate for this purpose.
6	Impact of new tool or process change	Difficult to assess with this approach as it would be necessary to divide up costs between different content streams that may have	The Lifecycle approach is appropriate for this purpose.

[2] As demonstrated by the Digitale Bewaring work at the Dutch National Archives, http://www.digitaleduurzaamheid.nl/bibliotheek/docs/CoDPv1.pdf
[3] As demonstrated by the LIFE approach to digital preservation costing, http://eprints.ucl.ac.uk/1854/

		different preservation processes associated with them.	
7	Comparison of analogue and digital preservation	Overall figures for an analogue and digital repository may enable some useful analysis. It is suspected that lower level information around specific collections or lifecycles will be more useful.	The LIFE² Newspaper Case Study has provided some very useful analysis in this area, suggesting this is an ideal approach for comparing analogue and digital materials.

3.4 LIFE Model v2

3.4.1 Introduction



The LIFE Model provides a view onto the typical processes applied to digital objects throughout their lifecycle by an organisation acting as the custodian of those objects. The processes are loosely organised in a chronological order, from their creation through to eventual access. It should be noted however that processes can, in practice, overlap with each other or be executed in a different order. The model aims to capture common processes found in most digital lifecycles. While some processes may not be applicable to all lifecycles, the intention is to provide meaningful placeholders for the majority of typical lifecycle processes.

3.4.2 Purpose of this Section

This section draws together feedback, discussion and review of the LIFE Model from a number of sources:

1. The LIFE[1] and LIFE[2] Project Teams, and the staff of their institutions
2. Feedback from review by an independent economics expert[4]
3. The LIFE[1] and LIFE[2] Project Conferences[5]
4. Early adopters of the LIFE Model (particularly the Royal Danish Library, State Archives and the State and University Library, Denmark)
5. Feedback from the LIFE[2] Case Studies

The result is a final update to the LIFE Model which builds on both the first published LIFE Model in 2006[6] and the updated LIFE Model v1.1 which was produced as a working update in 2007[7].

In line with the objectives of the LIFE[2] Project, this final revision aims to:

1. fix outstanding anomalies or omissions in the Model
2. scope and define the Model and its components more precisely
3. facilitate useful and repeatable mapping and costing of digital lifecycles.

[4] Björk, B.-C. (2007) *Economic evaluation of LIFE methodology*. Research report. LIFE Project, London, UK. Available from: http://eprints.ucl.ac.uk/7684/

[5] The LIFE[1] Project Conference, 20th April 2006, The British Library, London, http://www.life.ac.uk/1/conference.shtml
and the LIFE[2] Project Conference, 23rd June 2008, The British Library London, http://www.life.ac.uk/2/conference.shtml

[6] McLeod, R. , Wheatley, P. and Ayris, P. (2006) *Lifecycle information for e-literature: full report from the LIFE project.* Research report. LIFE Project, London, UK. Available from: http://eprints.ucl.ac.uk/1854/

[7] Wheatley, P., Ayris, P., Davies, R., Mcleod, R. and Shenton, H. (2007) *The LIFE Model v1.1.* Discussion paper. LIFE Project, London, UK. Available from: http://eprints.ucl.ac.uk/4831/

3.4.3 Feedback

The LIFE Team are very keen to continue to receive feedback and comment on this document, which can be directed to life@bl.uk.

3.4.4 Lifecycle Processes and Costs

Figure 3 outlines the stages of the latest version of the LIFE Model (v2). Figure 4 provides a more detailed view of the Model on a Stage and Element level.

Section 3.4.17 of this report provides notes and explanation on the justification for changes made. Section 3.4.8 provides a more detailed description and definition of the lifecycle elements.

Figure 3 - Stages of the LIFE Model v2

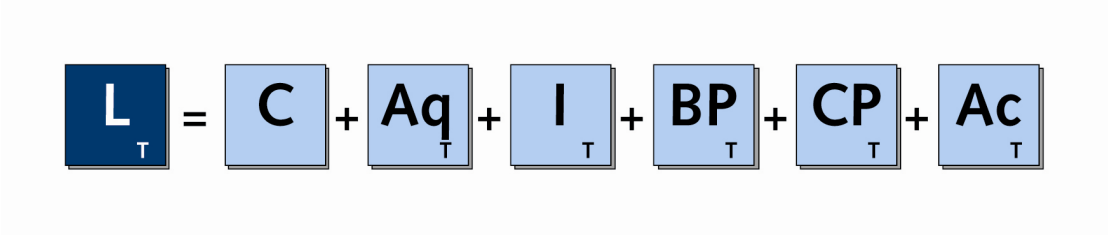


Figure 4 - The LIFE Model v2

Lifecycle Stage	Creation or Purchase [8]	Acquisition	Ingest	Bit-stream Preservation	Content Preservation	Access
Lifecycle Elements	....	Selection	Quality Assurance	Repository Administration	Preservation Watch	Access Provision
	....	Submission Agreement	Metadata	Storage Provision	Preservation Planning	Access Control
	....	IPR & Licensing	Deposit	Refreshment	Preservation Action	User Support
	....	Ordering & Invoicing	Holdings Update	Backup	Re-ingest	
	....	Obtaining	Reference Linking	Inspection	Disposal	
	....	Check-in				

[8] This stage may be beyond the scope of some costing activities. Creation may occur outside the view of the costing institution. It should therefore be considered to be optional. Where considered within scope, elements will need to be tailored to the specific lifecycle case in question.

3.4.5 Non-lifecycle Processes and Costs

The following Table provides a summary of non lifecycle costs. Scoping of lifecycle and non-lifecycle costs is discussed in more detail in the following section. For the purposes of the Case Studies the Non-lifecycle Stages were not considered, however, for institutions adopting the LIFE approach a decision on whether or not to include these costs needs to be made individually.

Table 2 - Non-lifecycle Stages

Non-Lifecycle Stage	Management and Administration	Systems / Infrastructure	Economic Adjustments
Non-Lifecycle Elements	Management	Repository Software	Inflation
	Administration		Discounting

3.4.6 Lifecycle Scope

3.4.6.1 Aims and Challenges of Defining the Scope of Lifecycle Costs

Defining the scope of "Lifecycle Costs" and "Non-lifecycle Costs" is essential if costing activities are to be precise and repeatable, and the results of costing activities are to remain comparable across different lifecycles and different institutions. Decisions in scoping must be practical, ensuring that costing activities are not unduly complicated. They should ideally reflect common sense in what should and should not be included and be consistent. Enabling a meaningful comparison between analogue and digital lifecycles requires careful consideration[9].

3.4.7 The LIFE Lifecycle Scope

Lifecycle Costs are considered to be the costs that are directly associated with the processes necessary to preserve some specific digital objects. The scope can be illustrated with the example of an established digital repository, with existing streams of digital objects coming into the repository. In this case, Lifecycle Costs could be considered to be the costs of whatever additional processes will need to be performed in order to add a new digital object stream to the digital repository.

Examples of Lifecycle costs: Selection policy, ingest of digital objects into a repository, creation or extraction of metadata, storage hardware, backup, provision of access.
Examples of Non-lifecycle costs: Management, inflation, digital repository software.
Hardware or software that supports a specific digital object stream is considered to be within the scope of lifecycle costing. Hardware or software that provides general support across all digital object streams is considered to be outside of the scope of lifecycle costing.

3.4.8 Stage and Element Definitions, with Suggested Sub-elements

Stages represent high level processes within the lifecycle which group related lifecycle processes together. Elements represent the next level down in lifecycle processes. They are still relatively high level and but are focused on a distinct process within the lifecycle. The

[9] Comparison between analogue and digital lifecycles is a key focus for LIFE², particularly with regard to the British Library Newspapers Case Study.

LIFE model attempts to describe a standard set of elements to which most digital lifecycles can easily be mapped. Sub-elements represent the specific components of a lifecycle element. At this level of detail, lifecycles are expected to vary considerably from one to another and so the detailed sub-elements are provided here for guidance only.

The breakdown of components within the LIFE Model is in Table 3.

Table 3 - Breakdown of Components for LIFE Model

Lifecycle level	Explanation
Lifecycle	The process from Creation to Access for a particular digital object, which can be broken down further into a number of distinct processes
Lifecycle Stage	A high level process within a lifecycle. Provides a way of grouping related lifecycle elements. Processes within a Lifecycle Stage typically occur or recur at the same point in time
Lifecycle Element	A distinct and significant lifecycle process that will provide useful costing information for organisations to perform planning, evaluative or comparative exercises
Lifecycle Sub-element	A suggested key component of a lifecycle element. Not significant enough to warrant inclusion as a distinct lifecycle element

3.4.9 Creation or Purchase

There are three main sources for digital objects which might be acquired and preserved by an organisation:

1. Creation (where the objects are created by or within view of the preserving organisation)
2. Purchase (where the objects are bought or licensed for use by the organisation)
3. Donation (where objects are donated to the preserving organisation at no cost)

This lifecycle stage is a placeholder for these different processes or costs that may be encountered by the preserving institution. Given the tremendous variations in cost that may be encountered that depend on the digital object stream in question, this lifecycle stage should be considered optional within the lifecycle costing process. It is therefore represented graphically within the LIFE Model separately from the other lifecycle stages.

Where a particular lifecycle involves creation within the preserving institution, particular lifecycle processes will need to be identified and defined. Examples of Creation processes might include: e-book authoring, scholarly publishing/authorship, and digitisation.

3.4.10 Acquisition

Acquisition represents the initial stages of acquiring and processing digital objects prior to ingest into a digital repository. Acquisition processes relate to collection management, administration and the operations of receiving or obtaining the objects themselves.

3.4.10.1 Selection

Selection is the key collection management process of deciding what materials should be acquired. This typically involves the development of a Collection Policy which will capture factors such as the mission of the organisation, the purpose and strengths of the collection and existing agreements influencing selection[10]. A selection process will then consider issues such as the value of the material, the expected use and expected costs of preservation against the drivers of the Collection Policy to decide whether to proceed with acquisition or not.

[10] Cedars Guide to Collection Management, Cedars Project, http://www.leeds.ac.uk/cedars/guideto/collmanagement/guidetocolman.pdf

Selection is typically conducted by a mix of collection management specialists, content specialists, operational staff and preservation staff.

The suggested sub-elements in Selection are listed in Table 4.

Table 4 - Suggested Sub-elements in Selection

Suggested Sub-elements	Explanation / notes
Selection Policy (policy/procedure)	Development of the Selection Policy for the collection
Selection (action)	The application of the selection process, guided by the Selection Policy
Selection Metadata (metadata)	Recording of metadata describing the scope, results and justification for the selection decisions

3.4.10.2 Submission Agreement

This is the process of establishing a submission agreement with the supplier of the digital objects being acquired. Requirements for the producers/depositors are established and agreement on the conditions of the submission will be negotiated with the producers/depositors. A submission agreement will define the details and conditions of the relationship between the acquiring organisation and the producers/depositor. This might include: the expected file formats of the digital objects, the packaging of the digital objects and expected medium of transport, the frequency of delivery of objects, and the procedures for mitigation should expected or agreed quality levels not be met.

For voluntarily- or legally-deposited digital objects, this element might focus on defining and communicating the conditions of deposit rather than specific communication and negotiation with the producer/depositor.

The suggested sub-elements in Submission Agreement are listed in Table 5.

Table 5 - Suggested Sub-elements in Submission Agreement

Suggested Sub-elements	Explanation / notes
Submission Agreement (policy/procedure)	Specification of submission requirements for producers/depositors
Negotiation of Submission (action)	Communication and negotiation with producers/depositors regarding submissions
Submission Metadata (documentation)	Recording of metadata relating to submission requirements

3.4.10.3 IPR & Licensing

IPR and Licensing is the process of researching, negotiating and agreeing on the rights to access and preserve digital objects. Research may be required in order to investigate the current IPR situation, and possibly identify and locate the rights holder. Negotiation with the rights holder may be necessary in order to agree on the right to access and preserve the digital objects. In some cases, rights may negotiated via a licensing agreement. It may be necessary repeatedly to re-negotiate agreements or to re-evaluate the IPR situation at particular times throughout the digital object's lifetime.

IPR and Licensing is related to the establishment of a Submission Agreement, but is considered to be significant enough to be treated as a specific lifecycle element.

The suggested sub-elements in IPR & Licensing are listed in Table 6.

Table 6 - Suggested Sub-elements in IPR & Licensing

Suggested Sub-elements	Explanation / notes
IPR and Licensing (policy/procedure)	This might include investigating the current IPR situation and who the relevant IPR holders are
Negotiation of Rights (action)	Negotiation of rights to preserve and provide access with producers/depositors
Negotiation of Licensing Agreements (action)	Negotiation of rights to provide access with producers/depositors
Rights Metadata (metadata)	Recording of rights metadata

3.4.10.4 Ordering and Invoicing

Ordering and Invoicing is the administrative process associated with ordering, invoicing and paying for digital objects, whether purchased or licensed. Following the establishment of a relationship with the producer/depositor as part of the Submission Agreement, and negotiation of rights, this element represents the more frequent and repeated communication to order, track and invoice for particular acquisitions or packages of acquired objects or titles. This may not be applicable for voluntarily or legally-deposited digital objects.

The suggested sub-elements in Ordering and Invoicing are listed in Table 7.

Table 7 - Suggested Sub-elements in Ordering and Invoicing

Suggested Sub-elements	Explanation / notes
Ordering and Re-ordering (action)	Ordering and re-ordering of the object, where it has been found to be of an insufficient level of quality during the Check-in or Quality Assurance processes
Invoicing (action)	Invoicing and administration for payments made
Ordering Metadata (metadata)	Record ordering and invoicing metadata

3.4.10.5 Obtaining

This is the process of transporting the digital object from the source via whatever means (for example by post on handheld media, by email, by ftp) to the preserving organisation. It is considered typical to utilise a checksum mechanism to guard against bit loss during the transport process, which would then be verified during the subsequent Check-in phase.

The suggested sub-elements in Obtaining are listed in Table 8.

Table 8 - Suggested Sub-elements in Obtaining

Suggested Sub-elements	Explanation / notes
Obtaining (action)	Transport of the object to the preserving organisation
Obtaining Metadata (metadata)	Record obtaining metadata

3.4.10.6 Check-in

Check-in is the process of ensuring that what was expected to be obtained (or ordered) 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 by matching against those that have been ordered. It might also include verification that bits have not been lost, by re-calculating and matching checksums.

The suggested sub-elements in Check-in are listed in Table 9.

Table 9 - Suggested Sub-elements in Check-in

Suggested Sub-elements	Explanation / notes
Content Check (action)	Verify titles, issues, filenames
Fixity Check (action)	Verify checksums
Check-in Metadata (metadata)	Record check-in metadata

3.4.11 Ingest

Ingest represents the processes involved in assessing and analysing digital objects and then ingesting them into the preserving organisation's digital repository.

3.4.11.1 Quality Assurance

Quality Assurance is the process of examining digital objects and ensuring they are of a sufficient or expected level of quality. If the assessed quality level is not sufficient, a mitigation strategy might have to be applied. This might include applying fixes, re-acquiring objects or recording metadata describing the details of the quality issues encountered. QA typically includes the process of checking the materials for viruses, and taking appropriate action to clean virus tainted objects.

The suggested sub-elements in Quality Assurance are listed in Table 10.

Table 10 - Suggested Sub-elements in Quality Assurance

Suggested Sub-elements	Explanation / notes
QA Policy (policy/procedure)	Description of quality requirements and required mitigation actions should quality requirements not be met. Policy for sampling of objects for QA (if appropriate)
QA Characterisation (action)	Characterisation of the digital object. Identification of file format, and assessment of whether the object is valid, well formed, and/or renders correctly with current access software
Content Examination (action)	Assessment of whether the content of the digital object is of an expected, agreed or sufficient level of quality. Typically, a manual process on a sample of the ingested objects
Mitigation (action)	Action to mitigate quality issues (might include virus cleaning or re-ordering or obtaining the digital object)
QA Metadata (metadata)	Record QA metadata

3.4.11.2 Metadata

This element represents the process of identifying, extracting and recording metadata describing the content. Characterisation is the process of examining content in order to understand its technical characteristics. Identification of the file formats of digital content is typically the first stage in this process. Further analysis can then be undertaken to assess the adherence of the content to the structures and standards of these formats (often referred to as "validation"). The extraction of metadata is the process of identifying and extracting metadata from the content and from existing metadata ingested with the content.

The suggested sub-elements in Characterisation and Metadata Extraction are listed in Table 11.

Table 11 - Suggested Sub-elements in Metadata

Suggested Sub-elements	Explanation / notes
File Format Identification (action)	Automated processing to identify the file format and related technical characteristics
File Format Validation and Integrity Check (action)	Automated matching of the content with the specifications of the format the content purports to be. May include verification that the content is valid and well formed. May include (sampled) manual checking that the content renders with the access software currently provided by the organisation or commonly used by their users
Metadata Extraction and Recording (metadata)	This sub-element is likely to be broken down further depending on the specific lifecycle in question. Metadata might be sourced and recorded from a number of sources. This might include metadata automatically extracted from existing metadata or from the content
Metadata Creation (metadata)	Create new metadata, typically as part of a manual cataloguing process
Record Event Metadata (characterisation)	A further breakdown of this sub-element should be made based on the implementing organisation's existing metadata structure. PREMIS and/or METS provide a guide to this breakdown

3.4.11.3 Deposit

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

The suggested sub-elements in Deposit are listed in Table 12.

Table 12 - Suggested Sub-elements in Deposit

Suggested Sub-elements	Explanation / notes
Deposit (action)	Commit the digital object to the repository
Deposit Metadata (metadata)	Record Deposit metadata

3.4.11.4 Holdings Update

This element refers to the updating of holdings records (e.g. catalogue) when new digital objects are accessioned.

The suggested sub-elements in Holdings Update are listed in Table 13.

Table 13 - Suggested Sub-elements in Holdings Update

Suggested Sub-elements	Explanation / notes
Holdings Update (action)	Update holdings records
Holdings Update Metadata (metadata)	Record Holdings Update metadata

3.4.11.5 Reference Linking

Reference Linking is the process of adding to or updating information used in systems that facilitate the finding of the digital objects.

The suggested sub-elements in Reference Linking are listed in Table 14.

Table 14 - Suggested Sub-elements in Reference Linking

Suggested Sub-elements	Explanation / notes
Create Search Indices (action)	Creation of indices for use within search engines
Reference linking (action)	Linking the object to entries in relevant finding aids
Reference Linking Metadata (metadata)	Record Reference Linking metadata

3.4.12 Bit-stream Preservation

Bit-stream Preservation is the process of storing and maintaining digital objects over time, ensuring that there is no loss or corruption of the bits making up those objects. Provision of storage on its own is not enough to constitute Bit-stream Preservation. Bit-stream Preservation can be achieved when storage is supported by effective management, backup, a programme of refreshment, and periodic inspection to ensure stored objects can be retrieved.

3.4.12.1 Repository Administration

Storage Administration represents general repository administration and other miscellaneous tasks associated with the provision of Bit-stream Preservation.

The suggested sub-elements in Repository Administration are listed in Table 15.

Table 15 - Suggested Sub-elements in Repository Administration

Suggested Sub-elements	Explanation / notes
System Technology Watch (action)	Monitoring for the need to upgrade or update systems or hardware due to technology obsolescence
System Security (action)	Maintenance and auditing of repository system security
Statistics and Reporting (action)	Recording and reporting of statistics
Disaster Recovery Planning (action)	Planning for recovery and re-establishment of the repository in the event of disaster[11]
Manage Duplicate Storage (action)	Management processes associated with effective maintenance and synchronisation of multiple node storage
Storage Procurement (action)	Procurement of storage hardware

3.4.12.2 Storage Provision

This element represents the process of storing digital objects, with the ability to retrieve them as requested[12]. It includes the support and maintenance of the storage hardware.

The suggested sub-elements in Storage Provision are listed in Table 16.

Table 16 - Suggested Sub-elements in Storage Provision

Suggested Sub-elements	Explanation / notes
Storage hardware (technology)	Costs associated with hardware purchases
Storage Maintenance and Support (action)	Maintenance and support necessary to keep the storage fully functional over time

[11] Note that this does not refer to Backup (covered in a subsequent element) or Duplication (covered in the next sub-element) as in the OAIS model.
[12] Note that multiple-node or multiple-site storage will require the modelling and costing of each node.

3.4.12.3 Refreshment

Refreshment is the process of moving stored items to new storage hardware as existing storage hardware reaches the end of its lifetime.

The suggested sub-element in Refreshment is listed in Table 17.

Table 17 - Suggested sub-element in Refreshment

Suggested Sub-elements	Explanation / notes
Refreshment (action)	Moving digital objects to new hardware

3.4.12.4 Backup

Backup is the process of making frequent copies of stored objects, typically on tape media, in order to provide a degree of insurance against lost, damaged or deleted data.

The suggested sub-elements in Backup are listed in Table 18.

Table 18 - Suggested Sub-element in Backup

Suggested Sub-elements	Explanation / notes
Backup Procedure (policy/procedure)	Development of backup policy and procedure
Backup (action)	Planned backup activity
Recovery (action)	Irregular (and hopefully infrequent) recovery of data from the backups

3.4.12.5 Inspection

Inspection is the process of ensuring stored objects can be retrieved without loss. It includes both automated fixity checking and manual retrieval and viewing.

The suggested sub-elements in Inspection are listed in Table 19.

Table 19 - Suggested Sub-element in Inspection

Suggested Sub-elements	Explanation / notes
Fixity Audit (action)	Automated auditing of stored objects ensuring matching re-generated checksums with previously stored checksums to identify changes or loss of content
Manual Inspection (action)	Manually inspection of a sample of digital objects to ensure they can be retrieved and rendered as expected
Inspection Metadata (metadata)	Record Inspection metadata

3.4.13 Content Preservation

3.4.13.1 Preservation Watch

Preservation Watch monitors the context in which the preservation lifecycle exists, and gathers requirements which will inform Preservation Planning activities[13]. These requirements will guide the decision process undertaken in Preservation Planning as to what action might be appropriate to take to preserve the digital objects.

The suggested sub-elements in Preservation Watch are listed in Table 20.

[13] And may also inform processes in other parts of the lifecycle.

Table 20 - Suggested Sub-elements in Preservation Watch

Suggested Sub-elements	Explanation / notes
Technology Watch (action)	Focusing on technology changes in areas such as file formats, rendering tools, environments
Monitor Institution (action)	Capturing preservation planning requirements from the preserving organisations preservation policy and broader organisational strategy
Monitor User Community (action)	Gathering requirements influenced by the end users of the objects
Monitor Producer (action)	Monitoring of the producer of the digital objects (if applicable)
Record Planning Requirements (metadata)	Recording of requirements for preservation planning based on information gathered by preservation watch activities

3.4.13.2 Preservation Planning

Preservation Planning considers inputs to the planning process such as the profile of the objects to be preserved, contextual factors such as usage, and other planning requirements (provided by Preservation Watch). It assesses available preservation solutions and develops a plan for preservation. A preservation plan should guide preservation staff in the actions required to preserve digital objects over time.

The suggested sub-elements in Preservation Planning are listed in Table 21.

Table 21 - Suggested Sub-elements in Preservation Planning

Suggested Sub-elements	Explanation / notes
Preservation Planning (action)	Assessment of planning requirements and preservation solutions, and development of preservation plans
Record/Update Preservation Metadata (metadata)	Updating preservation metadata, such as Representation Information, based on preservation planning conclusions

3.4.13.3 Preservation Action

Preservation Action covers the process of performing actions on digital objects in order to ensure their continued accessibility. It includes evaluation and quality assurance of actions, and the acquisition or implementation of software to facilitate the preservation actions. Preservation actions will be defined and described by a preservation plan created in the previous element.

The suggested sub-elements in Preservation Action are listed in Table 22.

Table 22 - Suggested Sub-elements in Preservation Action

Suggested Sub-elements	Explanation / notes
Integrate new preservation solution (action)	Obtain/integrate new preservation action tool
Perform Preservation Action (action)	Updating preservation metadata, such as Representation Information, based on preservation planning conclusions
QA Preservation Action (action)	Perform an evaluation and QA of the preservation action
Record Preservation Action Metadata (metadata)	Record Preservation Action metadata

3.4.13.4 Re-ingest

Re-ingest represents the ingest of migrated objects back into the repository. It might be modelled in different ways depending on the lifecycle and organisation in question, so is provided as a distinct element. Re-ingest might include the following repeated elements[14]:

1. Obtaining
2. Check-in
3. Quality Assurance
4. Characterisation and Metadata Extraction
5. Deposit
6. Holdings Update

It should also be noted that subsequently to Re-ingest, continued Bit-stream Preservation and Content Preservation of the new objects will also be required.

3.4.13.5 Disposal

Disposal represents the removal of digital objects from the repository if preservation is no longer needed. If a digital object is identified as unworthy of preservation then a disposal action will need to occur in accordance with a Disposal Procedure.

The suggested sub-elements in Disposal are listed in Table 23

Table 23 - Suggested Sub-elements in Disposal

Suggested Sub-elements	Explanation / notes
Appraisal Procedure (policy/procedure)	Development of appraisal policy and procedure to assess whether an object requires future preservation
Appraisal (action)	Appraisal process of a digital object to ascertain whether it should be preserved. If not, disposal should take place
Disposal Procedure (policy/procedure)	Development of disposal policy and procedure for permanent deletion of objects. As an object may contain sensitive information secure disposal needs to be guaranteed
Disposal (action)	Removal of all copies (access, preservation and back-up copies) of a digital object from the repository

3.4.14 Access

3.4.14.1 Access Provision

This element represents the process of providing access to the digital objects for users.

The suggested sub-elements in Access Provision are listed in Table 24.

Table 24 - Suggested Sub-elements in Access Provision

Suggested Sub-elements	Explanation / notes

[14] These repeated re-ingest elements may be more streamlined processes than when executed for the first time earlier in the lifecycle. For example, re-ingested content might be characterised to identify and validate file format, but more extensive metadata extraction may only duplicate what has already been captured and so might be omitted.

Access Provision (action)	Retrieval of digital objects and Provision of access to users
Rendering and representation (action)	Provision of software and/or information to facilitate rendering of the digital object by the user
Record Access metadata (metadata)	Record usage metadata

3.4.15Access Control

Access control represents the application of actions or technical measures to ensure access is provided to appropriate users as per previously-negotiated access rights, and that those users are only able to use the content in ways which conform to those rights.

The suggested sub-elements in Access Control are listed in Table 25.

Table 25 - Suggested Sub-elements in Access Control

Suggested Sub-elements	Explanation / notes
Access Control (action)	Restriction of access to those users allowed to use the digital objects
Technical Protection Measures (action)	Restriction in use of the digital objects
Record Access metadata (metadata)	Record usage metadata

3.4.15.1 User Support

This element represents the support provided to users who access the digital objects. It includes enquiry services, reference services and general user support and correspondence.

The suggested sub-element in User Support is listed in Table 26.

Table 26 - Suggested Sub-element in User Support

Suggested Sub-elements	Explanation / notes
User Support (action)	No further suggested breakdown below the element level

3.4.16 Non-lifecycle Processes and Costs

Non-lifecycle Processes and Costs are attributed to the preserving organisation but are not directly associated with the lifecycle of the digital objects in question. These might include staff management, administration (including financial and human resources), finance (including pension costs), facilities and their support (such as office space), and a range of economic factors (such as inflation and discounting). Costs that are not directly related to lifecycle processes should be considered as optional in terms of analysis and recording.

By isolating overhead costs from the costs directly related to lifecycle processes, the ability to compare different lifecycles (where one analysis accounts for overheads and the other does not) is retained.

In his review of the LIFE[1] Model and Methodology, Bo-Christer Bjork (2007) noted that in the MLA sector, it does not typically make sense to use discounting or to account for inflation[15]. However, there may be exceptions where these economic factors may need to be taken into account (for example, if a commercial company utilised the LIFE Model in costing digital preservation activity). Note that common trends in costs should still be covered in the lifecycle stages (e.g. increasing staff costs and decreasing storage hardware costs).

[15] Björk, B.-C. (2007) *Economic evaluation of LIFE methodology.* Research report. LIFE Project, London, UK. Available from: http://eprints.ucl.ac.uk/7684/

3.4.17 Changes to the Model

3.4.17.1 From LIFE Model v1 to v1.1

Figure 5 - Original LIFE Model v1

Lifecycle element	Acquisition	Ingest	Metadata	Access	Storage	Preservation
Element 1	Selection (Aq1)	QA (I1)	Characterisation (M1)	Reference linking (Ac1)	Bit-stream storage costs (S1)	Technology watch (P1)
Element 2	IPR (Aq2)	Deposit (I2)	Descriptive (M2)	User support (Ac2)		Preservation tool cost (P2)
Element 3	Licensing (Aq3)	Holdings update (I3)	Administrative (M3)	Access Mechanism (Ac3)		Preservation metadata (P3)
Element 4	Ordering and invoicing (Aq4)					Preservation action (P4)
Element 5	Obtaining					Quality assurance (P5)
Element 6	Check-in (Aq6)					

1. Creation or Purchase stage added. This was not of relevance to the case studies covered in LIFE[1] but is clearly going to be needed in areas such as the Burney Digitisation Case Study in LIFE[2]. It may be possible to provide sets of standard elements for different creation scenarios.
2. Acquisition has been expanded with an additional element: Submission Agreement. IPR and Licensing have been collapsed into one element.
3. Ingest remains relatively unchanged, other than the move of Reference Linking to here from Access.
4. Metadata has been renamed as Metadata Creation. Elements are now categorised by process rather than metadata type.
5. Access has been moved to the end of the lifecycle. While this is not a change of great significance, it gives preservation a greater emphasis with the implication that to achieve access preservation issues need to be addressed. Access Control has been added.
6. An attempt has been made to scope lifecycle and non-lifecycle costs, recording them in separate Tables. This is a difficult area and will require further attention and review.
7. Storage has been re-named as Bit-stream Preservation. Previously, the LIFE team had only one element here, and this required expansion. This stage has now been broken down into more specific elements.
8. Preservation has been re-named as Content Preservation. This is useful as the word Preservation on its own can be ambiguous.

9. The elements within (Content) Preservation have been refined. These are all subject to further change depending on developments to the Generic Preservation Model:
    a. Technology Watch has been changed to Preservation Watch to reflect the wider range of external entities and changes that need to be monitored.
    b. Preservation Tool Cost has been removed and is subsumed into Preservation Action.
    c. Preservation Planning has been added. It was previously covered by Technology Watch, but is seen to be important enough to warrant a specific element.
    d. QA is subsumed into Preservation Action, as an integral part of that process.
    e. Re-Ingest has been added. This is really a placeholder for repetition of Ingest (and other) elements following a migration action.

Figure 6 - LIFE Model v1.1

Lifecycle Stage	Creation or Purchase[1]	Acquisition	Ingest	Metadata Creation[2]	Bit-stream Preservation	Content Preservation	Access
Lifecycle Elements	. . .	Selection	Quality Assurance	Re-use Existing Metadata	Repository Administration	Preservation Watch	Access Provision
	. . .	Submission Agreement	Deposit	Metadata Creation	Storage Provision	Preservation Planning	Access Control
	. . .	IPR & Licensing	Holdings Update	Metadata Extraction	Refreshment	Preservation Action	User Support
	. . .	Ordering & Invoicing	Reference Linking		Backup	Re-ingest	
		Obtaining			Inspection		
		Check-in					

3.4.17.2 From LIFE Model v1.1 to v2

Following the publication of v1.1, the LIFE Model has been continually refined throughout the LIFE² project as a result of the feedback received during the project (outlined in the beginning of this section).

Key changes include:

1. Metadata Creation Stage removed
   The stage and elements have been incorporated throughout rest of model
2. Content Preservation Stage
   'Disposal' Element added

Figure 7 - Stages of the LIFE Model v2

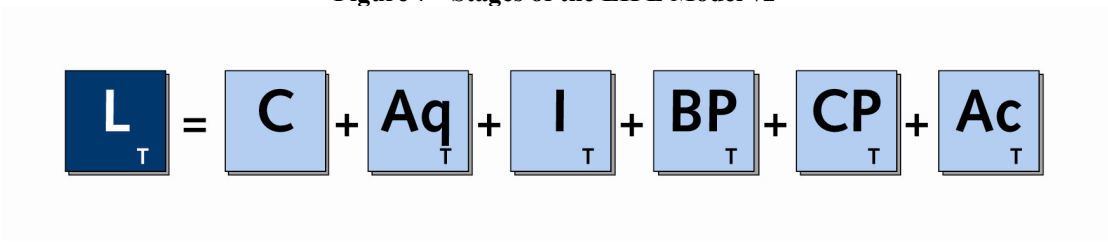


Figure 8 - The LIFE Model v2

Lifecycle Stage	Creation or Purchase [16]		Acquisition	Ingest	Bit-stream Preservation	Content Preservation	Access
Lifecycle Elements	....		Selection	Quality Assurance	Repository Administration	Preservation Watch	Access Provision
	....		Submission Agreement	Metadata	Storage Provision	Preservation Planning	Access Control
	....		IPR & Licensing	Deposit	Refreshment	Preservation Action	User Support
	....		Ordering & Invoicing	Holdings Update	Backup	Re-ingest	
	....		Obtaining	Reference Linking	Inspection	Disposal	
	....		Check-in				

3.4.18 Next Steps

A third stage of the LIFE Project would allow for the development of a fully-predictive toolset for LIFE. Based on feedback received at the LIFE² Conference from the HE, library and preservation communities, there is a clear need for the development of a LIFE tool.

Potentially it might also be worth looking at further development of case studies, as well as gaining additional feedback from the wider preservation community. Early adopters will also allow further development of the model. However, there are certain limitations to developing additional Case Studies, which are discussed further in the Findings & Conclusions section.

3.4.19 Glossary

Specific LIFE terminology used within this section is outlined below.

[16] This stage may be beyond the scope of some costing activities. Creation may occur outside the view of the costing institution. It should therefore be considered to be optional. Where considered within scope, elements will need to be tailored to the specific lifecycle case in question.

Table 27 - LIFE Terminology

LIFE Term	Definition / explanation
LIFE Model	The LIFE map or breakdown of a digital lifecycle (C Aq I BP CP A)
LIFE Methodology	The LIFE methodology for costing a digital lifecycle utilising the LIFE Model.
Stage	A significant process within the LIFE Model (e.g. Acquisition)
Element	A specific process within a LIFE Model Stage (e.g. Selection)
Sub-element	A specific process within a LIFE Model Element (e.g. Develop Selection Policy)

3. Models & Economics
Economic Review of LIFE
Digital Preservation Costing
LIFE Model v2
Generic Preservation Model

3.5 Generic Preservation Model v1.1

This section builds on the work done in the first phase of LIFE. Details of the original Generic Preservation Model can be found in the LIFE[1] Project Report (pages 90 to 107)[17].

3.5.1 Introduction

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[18]. 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 estimate the cost of 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.

3.5.2 Purpose

The purpose of this section is to outline a revision to the LIFE Generic Preservation Model (GPM) incorporating comment and feedback. This revision will be utilised in the first drafts of the LIFE² Case Studies. It is expected that a second revision, adding more detail, will be made after the completion of the LIFE² Project. This revision will include a review by an Experts Group, as originally outlined in the LIFE² Project Plan. Unfortunately the full review was not possible before project completion due to time constraints. What is outlined in this report reflects the first update by the Experts Group.

3.5.3 GPM Spreadsheet

The GPM is accompanied by an Excel spreadsheet which acts as a tool for estimating preservation costs using the GPM v1.1. The spreadsheet includes sample data from LIFE[1] which can be pasted over (formats, numbers of objects etc). Note that the POC (and other relevant constants) should be adjusted when applying the model.

3.5.4 Revised GPM

This section details the individual parts of the Generic Preservation Model.

3.5.4.1 Main Cost Elements

Technology Watch = *TEW*y*POC*

Technology Watch is the cost of monitoring tools, standards and other technology related to a particular format during the period of years costed (y). It is scaled by POC as the cost of Technology Watch does not need to be duplicated for each content stream at a particular institution.

Tool Setup Cost = *COA*ENP*POC*

Tool Setup Cost is the cost of preparing a particular tool so that it is ready to perform preservation actions. This does not include the cost of developing tools which is now considered to be outside the scope of a lifecycle cost in LIFE². This cost is based on a basic constant, the COA. Again, Tool Setup Cost is scaled by the POC.

---

[17] McLeod, R., Ayris, P. and Wheatley, P. (2006) "Lifecycle information for e-literature: full report from the LIFE Project", Available from: http://eprints.ucl.ac.uk/1854/

[18] See Cedars Project, Research Review, LIFE[1], http://eprints.ucl.ac.uk/1856/1/review.pdf

Preservation Planning *= (PLA*ENP+PLN*(y-ENP))*POC*FCX*
Preservation Planning is the cost of planning for preservation activities that will occur in the course of the costed period. The cost of Preservation Planning in a year when a Preservation Action will be executed will be high as more work is required. This base cost is defined by the PLA. The cost of Preservation Planning in a year when a Preservation Action will not be executed will be much lower as less work is required. This base cost is defined by the PLN.

Example:
Over a 10 year period (y=10) when it is expected that 1 Preservation Action will be conducted the costs will break down as follows:
(PLA * 1 + PLN * 9)

This base cost is then scaled by the POC and FCX.

Execute Preservation Action *= ENP*PON*n*(FCM/n+PUM)*
This is the cost of executing Preservation Actions during the costed period. The basic cost of migrating n objects is demonstrated in the table below, with savings made by economies of scale:

Table 28 - Cost of Migration

No. of objects to be migrated	PUM	FCM	Migration cost (total)	Migration cost per object
1	0.05	340	£340.05	£340.05
10	0.05	340	£340.50	£34.05
100	0.05	340	£345.00	£3.45
500	0.05	340	£365.00	£0.73
1,000	0.05	340	£390.00	£0.39
2,000	0.05	340	£440.00	£0.22
5,000	0.05	340	£590.00	£0.12
10,000	0.05	340	£840.00	£0.08
100,000	0.05	340	£5,340.00	£0.05
500,000	0.05	340	£25,340.00	£0.05
1,000,000	0.05	340	£50,340.00	£0.05
5,000,000	0.05	340	£250,340.00	£0.05

The basic cost as outlined above is then scaled by the Expected Number of Preservation Actions in the period (ENP). It is also scaled by the PON, which represents the percentage of the collection which will be preserved using a migration approach.

Quality Assurance *= BCT*n*ENP*FCX*
This is the cost of quality-assuring Preservation Actions conducted during the costed period. A base testing cost (BCT) is multiplied by the number of objects and is scaled by the complexity of the objects in question.

Expected Number of Preservation Actions (ENP) *= y/(BLE+0.1*y) + PON*
The Expected Number of Preservation Actions is the number of times it will be necessary to execute a preservation action during the period of years costed (for example in a 10-year

period this might average out to be 1.5 actions). Note that the PON considerably influences the Frequency. If a large number of objects will be preserved using normalisation, up front costs will be much higher.

3.5.4.2 Other Cost Elements

POC = Proportion of Collection (Generic tasks conducted as an organisation, such as Technology Watch, will actually support more than one collection. This factor represents the percentage of all content held at the institution in this particular collection or content stream. This enables a more accurate spread of the cost of generic tasks that effectively support more than one content stream).

PON = Proportion of Normalisation (This indicates the percentage of the collection or content stream that Preservation Actions will be applied to at the point of ingest)

POM = Proportion of Migration (This indicates the percentage of the collection that will be preserved using a migration approach. This impacts on the cost of physically-migrated content. Note that the GPM does not currently include re-ingest of migrated content)

FCX = Format Complexity (This provides a general indication of the complexity of a particular format which will impact on associated preservation activities. A low value represents a low complexity and a high value represents a high complexity. Examples are provided in the accompanying spreadsheet)

BCT = Base Cost of Testing (This provides a basic cost for testing or quality assurance activity associated with a preservation action applied to one digital object)

PLA = Planning: Action (The cost to perform preservation planning for a Preservation Action)

PLN = Planning: No Action (The cost to perform preservation planning in a year when no Preservation Action will be performed)

TEW = Technology Watch (The cost of a technology watch activity for a particular format)

BLE = Base Life Expectancy (The average time that a format will survive before requiring action to preserve it)

PUM = Per Unit Migration (The cost of migrating a single digital object)

FCM = Fixed Cost Migration (A one-off cost for migrating digital objects from one format to another)

COA = Cost of Action (The cost of setup and integration with preservation systems for a particular Preservation Action tool)

FSF = Format Stabilisation Factor (The annual change in file format life expectancy per-year)

3.6 Future Developments of the Models

Further development of Case Studies, as well as gaining additional feedback from the wider preservation community and early adopters would certainly allow for further development of the model. The LIFE team would be most interested in gaining further comment from institutions thinking of adopting the LIFE Model.

However, the LIFE team sees the next major step to further develop predictive models for the rest of the LIFE Model. A third stage of the LIFE Project would allow for the development of a fully predictive toolset for LIFE. Based on feedback received at the LIFE^2 Conference the HE, library and preservation communities, there is a clear need for the development of such a LIFE tool.

In terms of the Generic Preservation Model, the team will continue to work with the Experts Group setup at the end of this second phase to further discuss and refine the thinking around the Preservation Model.

3.7 Section Review

This section has briefly outlined the independent economic review commissioned by the LIFE Project (Section 3.2), summarised some of the key reasoning behind the need for costing digital preservation (Section 3.3), and outlined the current version of the LIFE Model (Section 3.4) and the Generic Preservation Model (Section 3.5).

In the next two sections we shall look at the Case Studies, in particular, how the LIFE Model was adopted for Institutional Repositories (Section 4) and for a comparison of lifecycle costs of both analogue and digital collections (Section 5). It is important to note that the LIFE Model used in these Case Studies was a working update to the Model (v1.1). The feedback from the Case Studies fed into a final update which led to the final revision of the Model (v2) which has been outlined in this section (3.4).

4 INSTITUTIONAL REPOSITORY CASE STUDIES

4.1 Purpose of this Section

Wishing to provide a Model that can be used throughout UK, and globally, to cost the lifecycle and long-term digital curation of deposited research outputs, LIFE[2] developed a range of costing studies to complement the outputs of the Case Studies in LIFE[1]. Based on repository development, using the SHERPA-LEAP and SHERPA DP Projects as testbeds for identifying lifecycle costs, this section analyses the Institutional Repository exemplar Case Studies used in the project.

These Case Studies, and therefore this section of the report, can be grouped into two areas – SHERPA DP and SHERPA-LEAP.

- **Supporting Documentation** lists what supporting documentation is available to support these Case Studies (primarily the spreadsheets) and how best to use them.
- **SHERPA DP Case Study** outlines the mapping of the repository services that CeRch provide to the LIFE Model.
- **SHERPA-LEAP Case Study** has been split into three repository Case Studies:
  - Goldsmiths, University of London
  - Royal Holloway, University of London
  - UCL (University College London)
- **Section Review** summaries some of the key findings from these Case Studies.

Each of the Case Study write-ups discusses in detail how the team used the LIFE Model, along with a brief analysis of the findings and any conclusions. Overall results and the conclusions that can be drawn from them are analysed in more detail in Sections 6 (Findings) and 7 (Conclusions).

4.2 Supporting Documentation

As suggested in the Introduction, before reading the Case Study write-ups, it would be best to have a general understanding of the LIFE Model which is outlined in Section 3. Each of the Case Studies is also supported by costing spreadsheets that are available for download from the LIFE website (www.life.ac.uk).

4.3 SHERPA DP Case Study

4.3.1 Background

The Centre for e-Research (CeRch) is a research department operating at King's College London that has been funded to develop the e-research infrastructure in King's and to contribute to the advancement of e-research in the national and international community. Prior to 1 April 2008, the department operated independently as the Arts and Humanities Data Service (AHDS) Executive, which co-ordinated a distributed service for the curation and preservation of digital collections produced by the arts and humanities research community. In this role, it collected research data produced by arts and humanities researchers, curated the data in a storage environment (with a current capacity of 15TB) and made it available for download. The AHDS Executive also participated in several research projects, investigating topics associated with long-term management of research data.

4.3.2 Introduction

The SHERPA DP project was funded by JISC and CURL (now RLUK) during 2005-2007. The AHDS worked with the University of Glasgow, University of Edinburgh, University of Nottingham, as well as the consortia of White Rose Research Online (Leeds, Sheffield, and York) and the London-based SHERPA-LEAP partnerships to develop a collaborative model and technical infrastructure that would address the requirements of each stage of the digital object lifecycle, from submission into a digital repository to transformation and subsequent use and re-use (http://www.dcc.ac.uk/docs/publications/DCCLifecycle.pdf). The collaborative model specified two types of institution – Content Providers and Service Providers – that had different responsibilities and performed different tasks in the lifecycle of a digital object. The five project partners, comprising a combined total of ten institutions, provided the infrastructure to accept research data submitted by academics and to make it available for use in the research community. The majority of project partners maintained the EPrints repository software, with the exception of the University of Edinburgh who operated a DSpace-based repository. During the funding period, the primary remit of project partners was to collect research papers written by academics within the respective institution. The research papers were mainly text-based, often accompanied by raster images that were embedded within the file. The Service Provider, a role performed by the AHDS Executive, was to take responsibility for the curation of digital objects made available by Content Providers, performing activities necessary to ensure the information contained in the research papers was accessible and authentic. The AHDS operated as a "dark archive" that performed curatorial action and returned replacement copies of the research papers to the Content Provider if the storage format was at risk of obsolescence. The technical infrastructure on which the project was based was created with two objectives: it should automate the process of transferring research materials from the project partners to the Service Provider, who could then perform curatorial action, as far as possible; and it should be atomic in its design, allowing software components to be replaced and/or extended as necessary. The AHDS was subsequently funded to extend the collaborative model in SHERPA DP2 to apply to a wider range of repository software and data types.

4.3.3 Relevance to LIFE work

The digital repositories operated by each institution in the SHERPA DP partnership are often funded to provide digital research materials for access and, in most cases, do not claim to be performing long-term curation and preservation. As a result, lifecycle costing of research materials stored by these repositories may be considered incomplete, lacking details of the

bit-stream and content preservation sections of the LIFE model. The research materials provided by partner institutions were ingested and curated in conformance with the Open Archival Information System reference model (ISO 14921: 2003) and Trusted Digital Repositories (TDR) specifications. Although costs for work were paid through project funding, the experience gained in the project provides evidence of the costing that may be applied to the development of a similar infrastructure for preservation services.

4.3.4 Scope of Case Study

The Case Study describes and costs the work required to establish the SHERPA DP preservation service, chiefly through development work enabling a Fedora-based archive to interact with the five chosen project partners, the majority of which were running EPrints (an exception is the Edinburgh Research Archive that maintains DSpace). The cost of storage and preservation are projected over a ten-year period. However, it has been assumed that no major work will be required to re-implement the data/metadata transfer mechanism. Subsequent to the initial implementation phase, the major costs are storage and preservation watch. No preservation actions have been assumed within this period.

4.3.5 Aim

The Case Study aims to show the costs relating to preservation when conducted as a third-party, or outsourced, service for IRs. It should be possible to compare them with the preservation costs for IRs undertaking preservation in-house.

4.3.6 A Note on Costs

All the staff costs are based on 2007/08 gross salaries to include National Insurance and superannuation costs. They exclude other overhead or oncosts: in the university sector, these oncosts are determined using the TRAC (Transparent Approach to Costing) methodology. At King's each full-time staff member has an Indirect costs overhead of £42,480 and Estates overhead of £12,998 (2007/08 rates). These costs are passed on in externally-funded research projects and other cases where the full economic costs have to be recovered. Inflation is assumed to be 4% per annum over the lifecycle, and the annual cost of capital (where indicated) is assumed to be 6%. Hardware and software are depreciated over four years in a straight line, making the annual rate 25%.

4.3.7 Mapping Case Study LIFE Model v1.1

4.3.7.1 Creation or Purchase

The project partners that participated in the SHERPA DP project offered research papers produced by academics in their institution as the target for their collection remit. A research paper (also called e-print) may be conceived, written and revised by one or more academics prior to its submission into an institutional repository. Many research papers authored in recent years are 'born digital', created using appropriate editing software and stored in file formats capable of storing text, images and metadata. In addition, a number of research papers previously existing only in printed form have been digitised by institution staff or the author at a later date to allow access and use. These documents may have initially been created by hand on paper, written on a typewriter, or created in an electronic form on a home computer and subsequently printed by the author.

The academic research papers may be deposited with institutional repositories on a voluntary basis or to comply with a department mandate to archive. A large number of the research papers, theses and dissertations are made available for free access and it is unlikely that the institutional repository has been required to make a payment to allow access and use of the

digital resources that they store. As a preservation service, the AHDS did not take a role in the creation or purchase of digital resources

4.3.7.2 Acquisition

4.3.7.2.1 Selection

The collection remit for each institutional repository that participated in the project specifies that they accept digital resources – draft papers that are being prepared for publication, peer-reviewed papers, dissertations and theses, and other textual resources, produced by academics operating in the institution. An assessment of the suitability of the deposited materials is made by institutional repository staff subsequent to submission. An institutional repository is likely to accept a large percentage of the research materials that have been submitted. However, a number of scenarios may be envisaged in which the institutional repository rejects or makes alterations to the content that has been submitted: it is written by author(s) employed by another institution; it contains content for which copyright belongs to a third party; the type of information is outside the collection remit of the institutional repository (e.g. an image collection).

The consortium agreement established between project partners indicates that the AHDS, in its capacity as a preservation service, is required to accept all digital resources that are made available by the Content Provider for harvesting. The Preservation Officer checks each year after the first that existing arrangements are still valid, taking one hour (£26) to complete this task.

4.3.7.2.2 Submission Agreement

An agreement is established between the depositor (an author or allocated third-party) and the institutional repository during the submission process to accept, store and make available research materials. The submission agreement for some institutional repositories omits the right to maintain digital resources in the long-term through migration or other preservation action. The AHDS has encouraged partner institutions to extend their submission agreement to define the rights required for preservation. The existing SHERPA consortium agreement was used as the basis for an agreement covering AHDS's preservation service; this took the Preservation Officer two hours (or £44).

4.3.7.2.3 Submission Metadata

The creation of resource discovery and administrative metadata is a key stage in the submission process. EPrints, DSpace and other repository software offer functionality for a depositor to submit metadata during the submission workflow, providing information on the creator, creation date, title of the paper, as well as other provenance and relationship descriptors. The responsibility for validating, enhancing, and/or correcting submitted metadata is allocated to an appropriate staff member in each institution. Resource discovery and administrative metadata are subsequently made available for review by researchers through a web interface and for machine-to-machine harvesting using oai_dc.

In the early stages of the SHERPA DP project, it was established that basic Dublin Core was insufficient and that the AHDS Preservation Service Provider required additional administrative metadata stored by the institutional repository, but not made available through OAI-PMH. The Preservation Officer performed a survey of the metadata stored by each institutional repository, including information that is not made available publicly (http://www.sherpadp.org.uk/documents/wp41-metadata_standards.pdf). Each repository stored fields that had been tailored to the unique requirements of the materials for which they were responsible. The Technical Officer modified the EPrints 2.x OAI export for each

repository to output further information stored in the underlying database as an OAI target, oai_dp. The AHDS Preservation Service harvested the oai_dp for the EPrints-based repositories on which work had been performed and oai_dc for the DSpace repository. The SHERPA DP project plan indicates that the above tasks required four weeks of Preservation Officer time to perform and write up the survey (£3,820), followed by 35 weeks of development work by the Technical Officer to implement the recommendations for four project partners (£39,846).

The metadata was harvested by, and files referenced in those records were transferred to, AHDS as automated processes. There were ingests at scheduled times throughout the project, but the system was designed to run monthly to capture new data.

4.3.7.2.4 IPR and Licensing

The rights for the institutional repository to accept, store, maintain and make available research materials produced by academic researchers are established in a submission agreement that is presented during the deposit process. Appropriate permissions for the AHDS Preservation Service to store and maintain access to research materials, on behalf of the institutional repository, is established

A contract between each institutional repository and the AHDS Preservation Service was written during the first three months of the project. The contract establishes permission for the AHDS to extract, store and maintain access to research materials made available on behalf of the institution for the lifetime of the project. The SHERPA DP contract is derived from an earlier contract written for the SHERPA consortium, which has been modified to include the activities specified in the project. It is estimated that the process of modifying an existing contract took two hours of the Project Director's time, which costs £62.

4.3.7.2.5 Ordering and Invoicing

One hour each year would be needed by the Administrative Officer to raise and process invoices relating to the service (£15). Since SHERPA DP was funded by a JISC grant, this cost has been discounted in the first year of the lifecycle.

4.3.7.2.6 Obtaining

The workflow developed for the project specifies that research data and metadata should be obtained at least once per month during the lifetime of the funding period. Several aspects of the project, including the obtaining process, are automated although there is the potential to manually initiate a data transfer process at any stage.

Each of the five content suppliers required some development work on their repository software to prepare for OAI-PMH harvesting at sufficient granularity. Repository staff were required to accept software code provided by the Technical Officer, as outlined in Submission Metadata, and to install it on a backup server for testing. The Technical Officer used the output to correct errors produced in the output and to provide a final version of the code for installation on the active repository. The work was funded by a grant from CURL for £25,000 – £5,000 per partner – as part of the joint funding arrangement with JISC.

4.3.7.2.7 Check-in

As with Obtaining, Check-in is an automated process which authenticates that requested data has been transmitted by comparing the URLs provided in the OAI target to the list of files that have been downloaded to the staging area. An error log is produced that specifies any

disparity between the two lists for review by the Technical Officer, taking an average of one hour (£22) per month. Subsequently, an automated fixity check is performed to ascertain that the data has not been intentionally or accidentally altered during the process of transferring data between the Content Provider and the Preservation Service Provider. During the project funding period (2005-2007), it was uncommon for a fixity to be made available in the OAI target, although it is increasingly common for repositories to make available an MD5 or other checksum calculation. The oai_dp target indicated the MD5 checksum generated on ingest by EPrints[19], which is used as the basis to compare and validate the transfer.

4.3.7.3 Ingest

4.3.7.3.1 Quality Assurance

The Quality Assurance process in the SHERPA DP ingest workflow is intended to automate the common activities involved in examining each digital object and validating its content, notifying an appropriate staff member (the Preservation Officer and Technical Officer) of any issues identified by the software, or other unexpected events that have occurred during the process.

4.3.7.3.2 QA Policy

The Quality Assurance (QA) policy created for the SHERPA DP project is derived from the requirement to maintain continued access to the information content, through a process of identification and minimisation of potential risks that will affect access. In the development of the QA policy two factors were considered to be important: the QA methods must be pragmatic in their implementation; and must be possible to perform using automated tools, with a minimal amount of effort for manual administration and error checking. The QA policy identifies three factors that must be considered: the identification of file format and version for subsequent obsolescence monitoring; identification of missing components in the object; and the presence of characteristics that are likely to impede preservation[20]. It is estimated that the process of designing the policy took two hours of the Preservation Officer's time at a cost of £52.

4.3.7.3.3 QA Characterisation

The characterisation of each digital object is a key component in the SHERPA DP workflow that is used to populate the preservation and technical metadata and to identify potential issues (as noted in the QA Policy). SHERPA DP utilises a combination of JHOVE and DROID to characterise the digital objects obtained from each institutional repository. The workflow[21] produced by the Technical Officer invokes DROID to perform an initial identification of the file format. If the format is identified (i.e. it is not an unknown bytestream) and is on a list of formats supported by JHOVE[22], JHOVE is invoked to provide a detailed analysis of the file format. The information generated during the aforementioned processes are used to populate the PREMIS and format-specific metadata for each record. In

[19] It should be noted that SHERPA DP2 validates data transfer using the MD5 checksum contained in the METS.

[20] see http://www.sherpadp.org.uk/documents/wp61-fileformats.pdf and http://www.sherpadp.org.uk/documents/wp65-migration_review.pdf for further information.

[21] A copy of the implementation specification that was used as the basis for the project may be downloaded at http://www.sherpadp.org.uk/documents/technical-specification.pdf.

[22] During the 2005-2007 period, the JHOVE 1 software recognised 12 distinct file formats and 45 encoding methods

the event that an error occurs (e.g. format mismatch between JHOVE and DROID DRM restrictions prevent a detailed analysis, etc.), a report is created and emailed to the Preservation Officer and Technical Officer.

4.3.7.3.4 Content Examination

The large number of digital objects provided by institutional repositories prevent a detailed analysis of the information content. It is presumed that the institutional repository that has provided the content will manually inspect the Information Content that it makes available for access. The QA workflow developed for SHERPA DP is intended to be entirely automated. However, some manual review may be necessary on a case-by-case basis, in the event that one or more errors occur. See mitigation (below in Section 4.3.7.3.5) for potential errors that may require the examination of the content.

4.3.7.3.5 Mitigation

The AHDS Preservation Service may perform mitigation in the event that an error is encountered in the processing of one or more digital objects.

Table 29 - Potential Errors and Required Mitigation Action

Error	Mitigation Action
Checksum is invalid	Reject the transferred file and re-initiate the process of obtaining one or more digital objects from the relevant institutional repository
Format mismatch between characterisation tools	Perform manual analysis of tool to establish cause of fault
Identification of encryption or other inhibitors	Contact institutional repository and request that a replacement is provided that does not contain the inhibitor. Accept data if a replacement cannot be provided

The primary cost of mitigation is the amount of time that must be allocated to resolving one or more issues by a staff member. Issue resolution may require a different amount of time that must be allocated on a case-by-case basis and, therefore, is difficult to estimate. On the basis of experience gathered in the two-year project, mitigation action takes an average of three hours and may occur approximately four times in a year. Therefore, it is estimated that the annual cost is £312 for 12 hours of Preservation Officer time for research papers. However, the amount of time allocated to the mitigation process may increase when handling a wide variety of material types that have increasing complexity. Further investigation into the area is being performed as part of SHERPA DP2.

4.3.7.3.6 QA Metadata

Technical metadata is automatically generated by JHOVE and/or DROID and massaged into the PREMIS data dictionary. The workflow produces log reports that indicate the automated actions that have taken place, including any errors that have occurred. However, the creation of provenance metadata for each activity could not be implemented in the project. Development in the area is ongoing as part of SOAPI and SHERPA DP2.

4.3.7.3.7 Deposit

In the SHERPA DP workflow, Deposit refers to the process of submitting a group of digital objects and metadata obtained from an institutional repository, as well as preservation metadata created during the characterisation process that is stored in a staging area and ingesting it into the Fedora implementation in use by the project. Fedora automatically creates an event record that indicates the date and time in which data has entered the system.

4.3.7.3.8 Holdings Update

The creation and update of holdings is automatically performed by Fedora on ingest into the repository.

4.3.7.3.9 Reference Linking

As a dark archive responsible for the curation of digital objects, it is considered outside the remit of the AHDS Preservation Service to create additional information that will support reference linking.

4.3.7.4 Metadata Creation

4.3.7.4.1 Re-Use Existing Metadata

The Fedora repository automatically populates a Dublin Core record for each object in the repository using the metadata bit-stream as a basis, as part of the ingest process. It is assumed that depositor-created metadata that has been validated and enhanced by the institutional repository offers an accurate description of each digital object[23].

4.3.7.4.2 Metadata Extraction

To curate and preserve the digital objects, the AHDS Preservation Service creates supplementary metadata that are likely to assist with maintaining long-term access to the information content. Preservation metadata conformant to the PREMIS Data Dictionary[24] and format-specific formats (TextMD for text documents, MIX for images) is created through the use of DROID and JHOVE.

4.3.7.4.3 Metadata Creation

Basic provenance information is recorded as PREMIS events, indicating the activities that have been necessary to obtain, validate, characterise and ingest each digital object into the Fedora preservation repository. Various scripts executed at various stages during the workflow produce most metadata created in the workflow. However, it is possible to create or edit event metadata manually if required. The latter has not been performed at the time of writing.

4.3.7.5 Bit-stream Preservation

4.3.7.5.1 Repository Administration

System Technology Watch

[23] Additional validation may be implemented at a later date, through integration of the MetaTools project outputs

[24] In the lifetime of the project, the PREMIS Data Dictionary 1.0 was the latest version available. Further work is being performed in SHERPA DP2 to refine and extend preservation metadata to support PREMIS 2.0 and above.

Technology watch is performed by a System Administrator as an ongoing process for all AHDS activities at the Centre for E-Research. Therefore, it is difficult to isolate the time allocated to the system on which SHERPA DP operates. The System Administrator allocates 10+ hours per month to the activity, of which one hour may reasonably be allocated to the project. The activity costs £180 as an annual rate.

In the initial stages of development, the Technical Officer undertook an assessment of storage requirements, investigated mechanisms for synchronising data between the source and target servers, and reviewed the use Fedora to support the preservation target of the IR content. These tasks took 16 weeks and cost £18,216.

System Security

The system in its entirety is maintained by the Systems Administrator, who identifies and patches potential system holes that may be exploited. The Technical Officer integrated tools for security, fixity and integrity into the Fedora workflow. The work took four weeks and cost £4,554. Ongoing monitoring by the System Administrator takes ten hours per month, of which one (costing £15) is attributed to the project.

Statistics and Reporting

The preservation system operated by the AHDS, similar to other types of network environments, produces a large number of reports on an ongoing basis. Reports created for SHERPA DP, including a record of objects ingested into the repository, are reviewed each month following the activity of transferring digital objects and metadata between the institutional repository and AHDS Preservation Service. The Preservation Officer takes an average of one hour (£26) per month to generate and analyse these reports. The System administrator monitors network performance throughout the working day while undertaking other tasks. The time allocated may equal one hour per week of System Administrator time, or £780 per annum.

Disaster Recovery Planning

This is covered in existing AHDS work, and so is not costed separately. Arrangements are in hand with an external service to provide data recovery, but this has never been used during the lifetime of AHDS.

Manage Duplicate Storage

The AHDS maintains a contract with the Science and Technology Facilities Council (STFC) to provide a mirrored copy of its digital collections, as well as materials provided by SHERPA DP partners, at an annual cost of approximately £31,000; a quarter of this cost is attributed to the SHERPA DP project, or approximately £7,750.

Storage Procurement

No additional storage was required during the lifetime of the project, due to the relatively small size of the materials provided by partners and the current development of existing AHDS infrastructure. It is estimated that staff time to procure new equipment in years five and nine would cost £2,070.

4.3.7.5.2 Storage Provision

Storage Hardware

The existing storage facilities in use by the AHDS are (currently) sufficient for the combined total of digital objects and metadata obtained from partner institutions. It is anticipated that the infrastructure (including storage) will require updating and extension every four years. A value of £55,000 has been placed on the existing infrastructure and, with a four-yearly replacement cycle, a quarter of this cost (£13750) has been allocated in years five and nine to cover the activity.

Storage Maintenance and Support

The KCL Centre for e-Research maintains a contract with the equipment supplier to cover maintenance issues, including disc performance monitoring, replacement on failure, and other support required to rectify problems. The System Administrator operates as the primary contact between the organisation and project/service-related support. It is difficult to extrapolate costs associated with SHERPA DP-related activities from the overall management of the operation. However, a Preservation Service must possess appropriate support contracts in order to operate. A proportion of the costs (£25,000 for a four-year agreement) has been added in years five and nine for the maintenance and support of infrastructure; as with storage hardware, a quarter of these costs – which relate to service contracts with the equipment suppliers – are assigned to the SHERPA DP project for lifecycle costing.

4.3.7.5.3 Refreshment

No refreshment of system has been undertaken recently, but planning will be needed before any significant further development can take place, as the system has nearly reached capacity.

4.3.7.5.4 Backup

Backup procedure

The AHDS has developed and implemented a backup procedure for the existing collections for which it is responsible. The task of developing backup procedures requires approximately two hours of Preservation Officer's time, equivalent to £52.

Backup

The AHDS utilises two backup methods that are performed each week:

1. The content of the preservation repository, including materials provided by partner institutions is automatically mirrored to an off-site location maintained by the STFC. These costs are covered by Manage Duplicate Storage above.
2. A manual tape backup of data is performed following the successful transfer and ingest of data and metadata, in concert with other AHDS backups. It takes approximately thirty minutes to select data for backup, two hours of machine time to write data to tape, and thirty minutes to validate the data. Tapes are subsequently transferred to an off-site location. To ensure continued use, new data tapes must be ordered four times per year (every three months) in a process that takes an hour. Backup activities cost £1248 (one hour a week or £1144 for backing up and £104 for tape ordering) annually, excluding the cost of the tapes.

Recovery

It has not been necessary for the AHDS to recover any data provided by partner institutions. However, experimentation in the area indicates that it will require approximately two hours of

System Administrator time to retrieve data stored on tape, or to transfer data from the off-site mirror.

4.3.7.5.5 Inspection

Fixity Audit

A monthly audit is performed as an automated procedure. The costs are included as part of the System technology watch.

Manual inspection

No manual inspection is undertaken for SHERPA DP, although this activity is applied to AHDS Collections. Depending on the object type, this might involve a word count of textual documents, checking that all columns in a dataset are present, and verifying that an image can be successfully opened.

Inspection metadata

The metadata from fixity audits is automatically generated.

4.3.7.6 Content preservation

4.3.7.6.1 Preservation watch

Technology Watch

Technology watch takes the Preservation Officer around one day per year for each format (£191, or £2674 for all 14 formats). This encompasses reading articles and news stories, checking specifications for new versions of established formats, analysing available files to see characteristics, and looking at access software. Environment scanning of this sort is intrinsic to AHDS activity as a data archive, so some activity may go unrecorded and the true annual figure may be greater than one day per file format. Since this activity is being undertaken as part of CeRch's remit, only one quarter of this cost is assigned to SHERPA DP (£6685).

Monitor Institutions

The Archive Manager and Preservation Officer each monitor King's College London for preservation planning purposes in running a digital archive. A small proportion of this activity is assigned to the SHERPA DP project, equivalent to one day per year for each member (£232 and £191 respectively).

Monitor User Community

The user community of the AHDS Preservation Service refers to the partner institutions that operate institutional repositories. The method in which the AHDS liaises with each partner varies according to their requirements. Potential activities may include monthly or bi-monthly email exchanges to identify any changes in requirements and the attendance of bi-annual meetings organised by the institution. For example, an institution may choose to switch from repository software A to software B. The monitoring of the user community was intrinsic to the development of SHERPA DP as a project and a figure of £3147 was assigned to travel expenditure. In addition, activity reports created by the AHDS for each partner each year require an hour of the Archive Manager's time at a cost of £31.

However, it is likely that a more formalised method would need to be adopted for a working service.

Monitor Producer

For the initial SHERPA DP project, the Producers refer to the academic researchers responsible for the creation of research data. The producers are likely to share common requirements, in terms of the method in which they access research materials. However, they may also have distinct needs that the institutional repository is better equipped to monitor.

Record Planning Requirements

Recording the process and updating the Preservation Handbook (see Preservation planning below) took the Preservation Officer around two hours (£52). No changes have been undertaken since harvesting and ingest began, and so there has been no subsequent need for record-keeping.

4.3.7.6.2 Preservation Planning

The AHDS undertakes preservation planning as part of its remit to curate digital research data in the arts and humanities. Research and recommendations on the handling of specific data types are distilled into a set of Preservation Handbooks[25], which fit into an overall management strategy, defined in a Centre Ingest Manual[26]. The Preservation Handbooks[27] were subsequently reviewed and re-written as two of the SHERPA DP work packages to consider the requirements of the partner institutions. The process of researching and revising the processing guidelines require approximately 10 days of the Preservation Officer's time in the first year (£1,910). One day per year is subsequently allocated for review and revision (£191).

Record/update preservation metadata

Preservation metadata are recorded in Submission Metadata above. It is envisaged that some revision of preservation metadata may be necessary, if an updated version of a preservation or technical metadata schema is released. For example, significant changes have been introduced in recent revisions to the PREMIS Data Dictionary and MIX metadata schema for still images, which will be assessed by the preservation manager and officer for any necessary work.

4.3.7.6.3 Preservation Action

Integrate New Preservation Solution

Due to the relatively short time period in which the SHERPA DP project was operating, it was unnecessary to implement a new preservation solution. The PDF/A specification was investigated, to identify if it was capable of containing significant properties and was

[25] http://www.ahds.ac.uk/preservation/ahds-preservation-documents.htm

[26] The Centre Ingest Manual is an internal document that indicates the workflow that must be followed to process research data, from negotiation to distribution.

[27] Knight, G. (2007). *An investigation of file formats in use by SHERPA DP repositories.* Available from: http://www.sherpadp.org.uk/documents/wp61-fileformats.pdf and
Knight, G. (2007) *Recommendations to ensure the long-term preservation of digital objects stored by institutional repositories.* Available from: http://www.sherpadp.org.uk/documents/wp65-migration_review.pdf

supported by necessary software tools. However, the specification was in draft and had only limited software support (Adobe Acrobat 7.1/7.2 provided basic support) and was not integrated into the system. The process required two days of work by the Preservation Officer at a cost of £382. For comparison, it has been identified that the integration of a tool into the preservation system requires five days of work by the Technical Officer, calculated at £920, and for the lifecycle an average of one tool every three years is costed on this basis.

Perform Preservation Action

The preservation strategy[28] developed by the AHDS establishes the file formats that will undergo bit-stream and content preservation; unknown and poorly-documented formats will receive bit-stream preservation only[29]. The ability to perform preservation action is affected by several factors, including the file format, type and complexity of the information content, and the availability and suitability of software tools. The data types provided by partner institutions, for which the AHDS were responsible, included PDF 1.1-1.6, JPEGs, HTML, text files and Corel Draw-structured diagrams[30]. To date, the AHDS has not performed large scale format migration actions. However, a small number of objects has been normalised to a preservation format: five JPEG images were normalised to uncompressed TIFF and re-ingested, which took the Preservation Officer under an hour (£26); similarly, seven Coral Draw objects were normalised to SVG. However, the conversion process for these failed to maintain the information content and the converted data was subsequently deleted. In the long-term, the AHDS is likely to normalise file formats considered to be at-risk, if a suitable preservation format and robust conversion tools are available. A similar amount of activity (one hour of Preservation Officer time, or £26) is projected for subsequent years, but the requirement and effort are both highly variable and subject to the content of the partner repositories.

QA Preservation Action

The preservation workflow developed for use by SHERPA DP is primarily an automated process. However, the TIFF and SVG objects referred to in 6.3.2 were visually inspected and compared. The format characterisation measurements were cross-matched, followed by a visual inspection to authenticate that the conversion produces the expected results. It took the Preservation Officer less than an hour to verify their successful migration (£26).

Record Preservation Action

Descriptors of actions and other events that occur during the time period in which each digital object is stored by the AHDS Preservation Service are recorded as PREMIS Events. Automated activities are recorded using standardised descriptions, which may be supplemented by human-created information if required. It is estimated that bespoke work will require four days per year of Preservation Officer time (£764).

4.3.7.6.4 Re-ingest

[28] Knight, G. (2007) *Recommendations to ensure the long-term preservation of digital objects stored by institutional repositories.* Available from: http://www.sherpadp.org.uk/documents/wp65-migration_review.pdf

[29] The collection remit of institutional repositories is, in most circumstances, limited to a small number of accepted file formats. However, the number of formats accepted is likely to increase if institutional repositories accept materials beyond e-prints.

[30] Knight, G. (2007) *Recommendations to ensure the long-term preservation of digital objects stored by institutional repositories.* Available from: http://www.sherpadp.org.uk/documents/wp65-migration_review.pdf

The AHDS Preservation Service may re-ingest objects in conformance to three scenarios:

1) The Preservation Service Provider establishes that digital objects obtained from a Content Provider have changed during transit, as indicated by differences in fixity and value. The data must be re-transferred, followed by activities necessary to validate and ingest the data.

2) The Content Provider publishes updates for one or more records and publishes them through SETS. The pre-defined workflow is followed on metadata harvest and data transfer, followed by additional activities to identify differences between versions of the metadata and data, and to import the updated information into an existing record held by the Preservation Service Provider.

3) The Preservation Service Provider creates derivative digital objects as a result of format conversion to a preservation and/or distribution format. The newly-created objects are ingested into the Fedora repository and appended to the relevant record.

4.3.7.7 Access

4.3.7.7.1 Access Provision

The SHERPA DP preservation repository operates as a 'dark archive' for institutional repositories. No user interface is currently provided to allow access to objects in the repository. The Technical Officer investigated means to offer web access for IR managers to metadata and objects stored at AHDS, but this was not implemented in the project. Development work for a prototype took three weeks and cost £3,415.

4.3.7.7.2 User Support

The users of the Preservation Service are identified as institutional repository staff in partner institutions. Support provided by the SHERPA DP project team included advice and guidance on the use of formats suitable for preservation and repository interoperability (e.g. software patches necessary to obtain a record in its entirety), frequent updates on development work and other report writing. The majority of project support was concentrated towards the end of the project during the implementation phase. User support required approximately one day per month of Preservation Officer time and half that for the Project Director (£191 and £116 respectively each month, or £3684 a year).

4.3.8 Difficulties in Mapping to Model

The Case Study relates to the SHERPA DP project, which models preservation services offered to institutional repositories by a third party. Development work created a harvesting mechanism for the automated ingest of IR content into an existing digital archive. As such, many of the costs of the archive are not expressed in the LIFE model; from recent work based on AHDS experience, it is expected that the annual staff cost of running a small data archive (3FTE) would be around £77,000. This figure and all those relating to staff costs are net of the indirect and estate costs applied in universities under the full Economic Costing (fEC) model – in the case of King's College London these are currently £42,480 and £12,994 respectively per FTE.

4.3.9 Comments on v1.1 of Model

Version 1.1 of the LIFE model separates out Metadata creation from other tasks associated with preservation. Metadata is intrinsic to preservation activity, generated or validated at different points in the lifecycle. No manual metadata was created in this project, but it would be difficult in the AHDS to separate out as a distinct workflow work that did create metadata.

4.3.10 Key Findings from SHERPA DP Case Study

Costs do not vary so greatly with quantities of data, for this process was largely automated. There were 6,526 objects harvested as part of the process for this project, and the costs are given in Table 30.

Table 30 - Summary of Costs from SHERPA DP Case Study

	Total cost £	Cost per object £	Annual cost per object £
Year 1	119,801	18.40	18.40
Years 1-5	317,711	48.70	9.70
Years 1-10	530,515	81.30	8.10

There were no costs for creation or purchase. Acquisition costs were mostly for the development of the OAI-PMH tool and integrating the harvester within the AHDS repository. Ingest costs were low, since quality assurance was the responsibility of the source repositories: scheduled harvesting using OAI-PMH led to file format characterisation automated using DROID. The largest cost area was in bit-stream preservation, since this included staff elements for system administration and technology monitoring, as well as provision for storage (including equipment renewal) and offsite duplicate storage.

Preservation action was the hardest part of Content preservation to cost, since it is cannot be predicted with great certainty. The team assumed a major task (two weeks' work) every three years to cover any such activity, but prior AHDS experience suggests such preservation action may be less likely. Other aspects of Content Preservation (Preservation Planning and Technology Watch) are more consistent across time – and of course, there were no re-ingest costs.

4.3.11 Conclusions from SHERPA DP Case Study

The SHERPA DP project aimed to determine how a third-party preservation service might work in practice, as a necessary step to designing such a service offered to others for a charge. The costing exercise in this Case Study determined the costs of the project, broken down into meaningful areas of operation. The only exception was in Metadata, which the team felt was not a separate cost area but intrinsic in different parts of the lifecycle. Even metadata characterisation was incorporated into the automated workflow as part of the acquisition process, and costed as part of the submission metadata element.

As a largely automated service, SHERPA DP could offer significant cost savings with increased quantity. During the next phase of SHERPA DP, the team will test larger ingest actions which will allow a new unit cost over time to be calculated, one which should validate this assertion. These new costs will help efforts to demonstrate the viability of a third-party preservation service.

More generally, the Case Study has given a meaningful structure to a costing exercise. The SHERPA DP activity mapped well to the LIFE model, even though it was significantly different from previous LIFE case studies in scope. Establishing costs has been of direct use in developing a cost model for third-party preservation, one which will help the team develop a business model to offer a charged service. In this respect, it dovetailed usefully with work on another Case Study (for the 'Keeping research data safe'[31] report) which looked at other repository and preservation costs in AHDS/CeRch.

Identifying costs at the activity level is a useful step in seeking greater efficiency through lowering costs or increasing throughput. In this Case Study, bit-stream preservation was identified as the major cost area. Therefore, the team needs to ensure that it has the cheapest-acceptable storage infrastructure, efficient system administration procedures, reporting mechanisms, etc.

It was not always straightforward to identify costs at the activity level for this Case Study, because of the distance of time for some activities and in unpicking elements of staff time assigned to SHERPA DP rather than other work. The effort was worthwhile, though, to gain a greater understanding of SHERPA DP's own costs and processes. It helps to have a business requirement for determining costs, but applying the LIFE model to different institutional settings is recommended to all with an interest in digital curation and preservation.

[31] Beagrie, N. *et al.* (2008) Keeping research data safe: a cost model and guidance for UK Universities. JISC.

4.4 SHERPA-LEAP Case Study

4.4.1 Introduction

This LIFE[2] Case Study has been prepared by representatives of three institutional members of the SHERPA-LEAP (London E-prints Access Project) consortium: Goldsmiths University of London, Royal Holloway University of London, and UCL (University College London). Each institution has prepared costings relating to the management of its institutional repository, using the LIFE Model v1.1, together with a commentary. Some shared observations on the model are also included (Section 4.4.7).

4.4.2 SHERPA-LEAP

SHERPA-LEAP (a partner in SHERPA) is a University of London (UoL) Consortium, led by UCL, which has helped to create open access institutional EPrints repositories at 13 University of London institutions. The SHERPA-LEAP web site can be found at http://www.sherpa-leap.ac.uk.

SHERPA-LEAP was established in February 2004 as a consortium of seven Higher Education institutions. All were then members of the federal University of London, whose Vice-Chancellor generously funded the project. The two over-riding aims of the project were to create EPrints repositories for each of the partner institutions, and to populate those repositories through collaborative advocacy. The repositories were hosted centrally by UCL. The seven development partners were:

- Birkbeck, University of London
- Imperial College London
- King's College London
- London School of Economics and Political Science (LSE)
- Royal Holloway, University of London
- School of Oriental and African Studies (SOAS), University of London
- UCL (University College London)

In 2005, the Vice-Chancellor generously awarded funding for a second, 18-month, phase of SHERPA-LEAP. The purpose of the second phase was to extend the partnership by inviting more institutions in the federal University of London to join the SHERPA-LEAP consortium, and to provide support for the creation and maintenance of EPrints repositories at every UoL institution, regardless of platform. This second slice of funding enabled the appointment of a full-time SHERPA-LEAP Project Officer.

13 UoL institutions are currently partners in SHERPA-LEAP. The following institutions joined SHERPA-LEAP during its second phase:

- Goldsmiths, University of London
- Queen Mary, University of London
- The School of Pharmacy, University of London
- School of Advanced Study, University of London
- The Institute of Cancer Research
- Institute of Education, University of London

A third phase of SHERPA-LEAP was funded to run until the end of July 2008. This has enabled the consortium to continue to offer support for EPrints repositories within London. A cross-searching service for the repositories has also been developed during this phase.

SHERPA-LEAP was created to move forward the repositories agenda in London. A UCL-hosted repository is not a condition of SHERPA-LEAP membership: membership is open to any institution from within the UoL with an EPrints repository, or with plans to develop one. As many as nine institutions have at one time been part of the UCL-hosted repository service. A number of the earlier partners, including LSE, King's and Imperial, took advantage of the SHERPA-LEAP service and network to identify and review their wider institutional requirements for a repository, before migrating to local platforms. Several partners are considering the future of their repositories beyond the lifetime of the hosted service. Nonetheless, the repositories continue to grow, with new content being added daily. Moreover, download data compiled by the consortium shows that the content of the repositories is being heavily used, with some reposited resources commonly receiving well over 100 downloads per month.

Within the partnership there is substantial diversity: the partner institutions represent a mixture of size and mission, ranging from the large, multi-disciplinary and research-led institutions, to the smaller and highly-specialised bodies. Three differently-sized and differently-orientated institutions participated in the LIFE^2 Case Study.

4.4.3 Scope and Limitations

The costings in the Case Studies have been limited to the costs relating directly to repository management. These may include staffing, software, and line management costs, and may include both capital and recurrent costs; but wider costs such as those of institutional content creation, the repository's share of the University communications and estates infrastructure, and management costs (other than those applying within the repository unit) are out of the scope of the studies.

4.4.4 Case Study 1: Goldsmiths, University of London

4.4.4.1 Background

An institutional repository at Goldsmiths, University of London was set up in 2006 and has been live since January 2007. The repository uses the EPrints software. Hosting and technical support were provided throughout the first two years by SHERPA-LEAP.

Goldsmiths Research Online, the institutional repository, was planned in consultation with the Research Office at Goldsmiths. It was intended to represent the diverse, creative qualities of the research environment at Goldsmiths. Journal articles, conference papers and presentations form a high proportion of the repository content. Events and activities such as concerts, conferences, talks, performances, screenings, exhibitions, artist residencies and online CVs are also represented as is a sample collection of practice-based theses in fine art.

During the first two years the repository was run as a 'test-case'. There were four planned phases of development: an initial pilot project provided a mediated deposition service for academics in the Psychology Department, initially 4 articles each, which resulted in the reposit of journal articles. The second phase, beginning in June 2007, was an open invitation to all academic staff to deposit scholarly work in the repository. There was a very low response to this. The third phase, running simultaneously, was to work in consultation with the Art Department and targeted individuals to focus on identifying how visual practice-based research was currently represented on departmental web pages and how the repository could

also present this kind of research. The aim during this phase was to develop procedures for dealing with the types of digital objects stored online. The fourth phase extended the consultation process to the Design and Visual Cultures Departments in late 2007. Data and documents collected during the RAE2008 process would be imported into the repository as a basis for the departmental collections. February 2008 then brought a review of the project in order to plan for the future development of an institutional repository at Goldsmiths, one that would extend beyond the end of SHERPA-LEAP in July 2008.

4.4.4.2 Relevance to LIFE work

The repository at Goldsmiths has been included as a Case Study since it contains examples of research output from the visual and performing arts disciplines. Few repositories hold this kind of material.

4.4.4.3 Scope of the Case Study

The costs of the first active year (2007) were assessed for the LIFE[2] Case Study. The repository remained relatively small (250 items), but included a variety of research types and formats, reflecting the focus on receiving outputs from art and design subject areas in this period.

The time spent by repository staff on the various elements of the LIFE model was estimated retrospectively and then assessed for the Case Study. Costs for each type of output (based on output types used in RAE2008 + theses) were calculated separately on the basis of repository staff time spent per item, then an average cost was produced for the records/objects that had been acquired over the preceding year. The actual figures quoted should be regarded as estimates, and the Case Study itself as illustrative.

4.4.4.4 Aims

Goldsmiths utilised the LIFE model to assist in the evaluation of the development of the repository as a 'test-case' project. The institution sought to identify the problematic elements of including practice-based arts research in repositories, based on a small number of representative examples. The aim was to produce an indication of the costs of including each specific type of research output in the repository. Goldsmiths also wanted to use the model to assist in assessing, again for each specific type of output, whether collection, cataloguing and curation fell within the current scope of the repository. The model was not used to make any comparison between analogue and digital lifecycles.

4.4.4.5 Comments on Lifecycle Processes and Costs

4.4.4.5.1 Non-lifecycle Processes and Costs

Comments on management, staffing and software are included in Section 4.4.6.9 on page 72. The following comment is specific to Goldsmiths.

4.4.4.5.2 Staffing

Various staffing permutations, involving portions of the time of permanent staff taken from other duties, have been in place since the repository was inaugurated. The role of repository administrator was incorporated into the job of Research Support Librarian (then graded AR3) in September 2007 as 0.3 FTE of that post. Various clerical assistants (graded CR3) worked on the repository for 2 hours per week for approximately 6 months over the total period (mostly in the summer vacation).

The Research Support Librarian was involved in planning and managing the repository as a whole and carried out the policy development elements of the lifecycle. She oversaw the day-

to-day management of the repository, was responsible for advocacy and liaison with Goldsmiths academics and line-managed clerical assistants assigned to the project. In addition, during this test stage, the repository administrator carried out all life-cycle tasks for the item types: article (though not for all articles), book, book section, conference, database, exhibition, monograph, other (comprising output from research projects), performance, software, thesis and visual/digital work. The clerical assistant carried out the IPR, metadata, obtaining, check-in and ingest stages only, of a scholarly work's lifecycle for the majority of articles and for all internet publication, only.

This model was not able to support the service at a level which matched potential deposit and the Library has recently been successful in securing funding for a one-year clerical assistant at 0.5 FTE.

For the purposes of this report, costs were calculated per item, using an hourly pay/salary rate for AR3 and CR3 posts. Time spent was roughly equivalent overall to 0.3 FTE of the Research Support Librarian's post and 0.1 FTE of a clerical assistant post. This is a reasonably close approximation of the staffing resources which the Goldsmiths Library was able to invest in the repository during the twelve months preceding the preparation of this study.

4.4.4.5.3 Repository software

The LIFE model states hardware or software that provides general support across all digital object streams is considered to be outside the scope of lifecycle costing. However, it should be noted that open-source repository software is evolving and repository administrators will be expected to upgrade repository software as new features/functionality are developed. For example: EPrints v.2 was installed at Goldsmiths and customized by adding all RAE output types to the printed types initially available. One year later, the software was upgraded to EPrints v.3, which included these types as standard. The upgrade itself presented problems in the transfer of metadata from one version to the next, and for which solutions were found and implemented. However, ensuring the sustainability of the resource could be considered a lifecycle cost.

4.4.4.5.4 Overall approach

The costings were based on the time taken to produce the 250 records, then in the repository. The repository was run as a 'test-case.' The repository administrator was involved in all the planning and administration of the repository, having supervised all aspects of the lifecycle. None of these factors was known at the start of the project. In an adapted spreadsheet, time spent on each element of the lifecycle was recorded in minutes, this was then multiplied by the per-minute pay/salary rate of staff time, and this figure was then multiplied by the number of items of that type added to the repository during the period. The costs were entered in the standard spreadsheet. Management and administration were estimated, and divided to produce a figure per item where possible. It was also decided to use the data relating to digital and visual media only in the LIFE report because it was found that the costings for different types of research output were very different, and fewer institutions have included this type of material in their repositories, thus providing a unique contribution.

4.4.4.6 Lifecycle Costs: Commentary

4.4.4.6.1 Lifecycle Processes and Costs

As the aim was to produce an indication of the costs involved with including each specific type of research output in the repository, costings were calculated per item. As there is a wide variation in costs between different types of item (from £7.93 per item for an internet publication, to £165.88 for a database), a separate lifecycle costing was produced for each item type. Attached in the appendices is the lifecycle costing for the 'visual/digital' type of item which included videos and documentation for digital works.

4.4.4.6.2 Creation or Purchase

The costs of content creation are in general out of scope for this study. In a university setting, content is usually donated. However, in repositories for the creative and performing arts, what may happen in effect, in artistic terms, is that a new, site-specific digital object is created for the repository which represents the art-work. The artist may want to be involved in how their work looks and feels in this new context and the repository administrator may be involved in negotiation about these aspects.

Concerning 'Visual/digital' item types, some of these were videos, represented by digital video clips. Depositors had created these as full-quality DV-PAL files. The £0.50 item creation cost shown in the lifecycle spreadsheet represents the time taken by the repository staff to communicate with the depositor explaining that the team would like a clip in a web-friendly format, asking whether the depositor would like to provide this or whether the repository staff should do so, and getting approval of the final version, all of which the team estimated taking an average of 1 minute per item. Similarly, the remainder of the visual/digital items ingested in the first year consisted of documentation describing the research project producing digital works. These were initially deposited in a PowerPoint presentation. Again, the £0.50 item cost represents an exchange between the researcher and repository staff member to clarify what the work consisted of, asking for supplementary information and covering the initial re-formatting of content of the original digital object.

It should be noted that as objects are created by or within the view of, the preserving organisation the repository administrator may advise on the use of standards at the planning stage, the selection of metadata standards for research databases, or formats for the deposit of videos, and lastly may need to negotiate how research is re-presented in the repository. It is likely that this advisory role will increase in the future as repositories become responsible for the curation and preservation of digital assets. The Versions Project[32] has produced helpful guides for depositors.

4.4.4.6.3 Acquisition

Deposit within the repository is voluntary and mediated. Researchers are encouraged to deposit research outputs after publication. In the subject areas represented at Goldsmiths, there is no tradition of the circulation of pre-print articles, and the team found that the majority of publishers of journal articles allow the post-print rather than the pre-print versions to be stored in institutional repositories.

Deposits are sent by email or in the case of larger files (videos and databases) posted on DVD or CD. Deposits may be sent as attachments to a generic email address or to the repository administrator. Researchers were given the option to self-archive, that is to upload metadata and objects directly to the repository, but this generally was not taken up.

[32] Versions Project website: http://www.lse.ac.uk/library/versions/

Additional metadata is requested from the depositor by the repository administrator via email. This is often necessary when notes on context and versioning have not been provided by the depositor.

4.4.4.6.4 Selection

There has been continuous re-evaluation of the collection policy, not only in terms of what is eligible content but on why content should be included and for what purpose. This is due to the wide variety of types and formats of the research output presented. The cost of selection (per item) is therefore calculated as a proportion of total time spent in meetings and in writing and revising the acquisitions policy.

Any Goldsmiths-authored, -created, co-authored or co-created research output is eligible for inclusion in the repository. Digital objects in a variety of types and formats, such as text, video, sound and image files, are stored.

Table 31 - File types in Goldsmiths Research Online (December 2007)

File type	Extension	Number of objects
PDF	.pdf	150
QuickTime movie, MPEG-4	.mov	7
JPEG image	.jpg	30
PowerPoint slideshow (containing images and video clips)	.pps	10
HTML file	.html	75
Sound file	.mp3	1
TOTAL		**273**

It should be noted that some types of research outputs (databases, software, performance, exhibitions) describing the material are represented in the archive by documentation, rather than by the full-scale output itself. This is for several reasons: the ephemeral quality of some practice-based research, IPR restrictions, practical or technical limitations, or conditions of production in specific milieu.

Supplementary data is accepted but may, however, be stored in another digital collection preserved by Goldsmiths, such as the Library's CALM ALM database or via the web-streaming service. In such cases, appropriate navigation is put in place between publication(s) and data. Where supplementary data is available and preserved in existing publicly-available sources (e.g. Visual Arts Data Service, or BioMed Central), links are provided to these.

Submission Agreement

The submission agreement should be reviewed periodically. While a default agreement is included in the default EPrints software, Goldsmiths feel this has become outdated as institutional requirements change over time.

Whilst the current LIFE model does not include advocacy as a lifecycle element, and advocacy has consisted partly of attending meetings to give reports and discussing departmental participation using various communication channels in Goldsmiths, the kinds of activity which could be absorbed into the management costs include the one-to-one meetings with individual academics and departmental technical staff to discuss how to represent or reformat research outputs.

IPR & Licensing

IPR discussions and negotiations take a substantial amount of time in the arts and humanities disciplines, as both artists and book publishers must be contacted individually. This has been less a case of making formal agreements or licenses than of discussing how to cover mutual requirements and establishing trust. The element 'negotiation of rights' was calculated as an overall figure, rather than per item, and then divided by the number of items included.

Check-in

Receipt of deposits was confirmed by email, which was then filed. Cost was calculated on the basis of average time spent per item. Content checks were carried out. No fixity checks were undertaken.

4.4.4.6.5 Ingest

Quality Assurance

The digital object is, in most cases, converted to a format suitable for curation and delivery on the web before deposit takes place. The first stage of 'Content Examination' is relatively lengthy for visual/digital items. Visual/digital items require 10 minutes of staff time per item and each of the 13 items of this type required individual assessment. In contrast, item types such as an article or book chapter did not require this kind of examination and required only 1 minute of staff time, despite needing reformatting. Text documents were converted to PDFs, as were most PowerPoint presentations and image slideshows. Videos needed to be compressed into appropriate codec and file types. Formats that can be viewed using freely-available software are preferred. Guidance regarding standards used for image files was sought from TASI, and the recent extension of their remit to cover sound and moving image formats will be useful to those who include these materials in their repositories.

In the subject areas covered (arts, humanities and social sciences), it was found that academics do not customarily keep an author's copy of the post-print version of published articles. The cost of formatting a version for the repository was therefore significant. During this first year, Word documents were reformatted and tagged using a standard stylesheet and a cover sheet was produced manually, for each article, before documents were converted to PDF.

The most complex digital objects were potentially the most expensive to ingest. Eight items described in the repository as the following research types, database (£38), software (£93), performance (£38), thesis (£48) and 'other' (research projects in most cases) (£83), cost on average eight times as much to ingest as articles (£20.10), books (£8) and internet publications (£3.50).

The requirement to develop the repository to represent practice-based research in the visual arts in the second phase had a significant effect on the costing. Further work needs to be done in this area as problematic issues remain unresolved: standardisation may have an effect on the rendering of the output that may be unacceptable to the creator, and standards used to ensure long-term preservation are different to those customarily used in repositories focussed on web delivery.

While it might be expected that importing existing research outputs would produce a lower unit cost, updates to software or charges for hosting and technical support for necessary

customization will be recurrent, lifecycle costs. The technical and IPR challenges raised when working with complex multi-media research outputs indicate that costs could continue to rise.

Lifecycle items Policy and Characterisation were estimated overall, at 2 hours and 4 hours respectively and an average unit cost calculation was calculated from these figures. Characterisation, particularly, was difficult to assess as standards for various types of digital object were investigated over a period of time.

Holdings Update

Holdings and associated metadata are updated when additional versions of existing digital objects are accessioned. Comments about new versions that replace older outputs are recorded in the repository. When an additional object is a version at the level of expression rather than an updated version, this is recorded as a note in the metadata.

4.4.4.6.6 Metadata Creation

Metadata creation takes place during the deposit stage of the life-cycle, preceding content examination. On receipt of a research output, repository staffs check the relevant copyright permissions, negotiate with publishers when appropriate, and create a full metadata record. Peer-reviewed items are distinguished from other deposits, and items 'in press' are marked as such.

Metadata-only records are accepted, typically where an author or department is keen that the repository should be able to supply a complete publications listing, or where copyright permissions prevent a full publication from being added to the repository. Less than 5% of records in the repository are currently metadata only, but it is anticipated that this will increase when those records that had been submitted to the RAE2008 are ingested.

Every new record is initially placed in a holding area for a quality assurance check by the repository administrator Typical checks include the copyright status of the research output in question, the bibliographic accuracy of the metadata, the inclusion of any external links and identifiers, and the integrity of the digital object.

Metadata for standard bibliographic resources is created by the clerical assistant but the cost of cataloguing visual and digital works, performances, databases and other non-standard works increased as they were produced by the repository manager. Once procedures were established and consistent staffing was in place, this task would be transferred to the clerical assistant.

4.4.4.6.7 Bit-stream Preservation

A Digital Preservation Policy for Goldsmiths is planned. For the current project, strategies for bit-level preservation are dealt with by UCL as the current host. Long-term preservation of complex digital resources has not yet been provided for but is being assessed by the SHERPA DP2 project funded by JISC, to which Goldsmiths is a contributor. It is likely that an outsourced model would be used in the future.

The costs associated with hardware are not applicable in the first year costings as SHERPA-LEAP absorbed the costs associated with the UCL-hosted repository service.

4.4.4.6.8 Content preservation

Preservation Watch

Preservation watch is carried out as a continuous process. The repository is still young, and formats utilised are known to meet current standards or best practice in respective fields, such as the work of the AHDS Subject Centres and TASI. It is desirable that repository staff attend relevant seminars and training sessions on content standards, which then presents another recurrent staffing cost. In the first year, one staff member attended a BUFVC course entitled Encoding Digital Video for Streaming and Network Delivery – Introduction (£153), and another attended the TASI course on Images and metadata (£150). One staff member attended a virtual training session on Adobe Acrobat (free), whilst two staff members attended a half-day course on new features in Adobe Acrobat 8 including PDF/A (funded by SHERPA-LEAP). An estimated £250 p.a. would be required for staff training on suitable formats.

Preservation Planning

Planning activities were estimated overall at 1 hour but were included as a proportion of the average unit cost. However, preservation issues were not considered comprehensively in this project.

4.4.4.6.9 Access

Access provision entailed selecting the correct digital object type from the list provided in EPrints, and thereafter the repository viewing software managed the creation of thumbnails and previews. This was estimated at 1 minute per item. Rendering and representation is defined as the provision of information to facilitate rendering of the digital object by the user. This is included to a limited extent in the repository software. However, rendering and representation also entailed selecting images to accompany videos as 'poster' images, and the organisation of an array of digital objects to represent an output, with appropriate version metadata. This element was therefore costed at £2.50 per visual/digital item, indicating 5 minutes work.

Access Control

In rare cases (e.g. embargoed theses or journal articles) the metadata profile of individual records is altered and access control is automated thereafter.

User Support

This is carried out as part of repository administration duties. Contact from users has so far been minimal, and email is the form of communication used.

4.4.5 Case Study 2: Royal Holloway, University of London

4.4.5.1 Background

4. Institutional Repository Case Studies
SHERPA DP Case Study
SHERPA DP Mapping
SHERPA-LEAP Case Study
Goldsmiths
Royal Holloway
UCL

An Open Access repository for Royal Holloway, University of London had been under discussion for several years before 2004, and the Liaison Librarian for Biological and Earth Sciences, Adrian Machiraju, had attended several meetings on the topic. Hence the announcement of SHERPA-LEAP was welcome, and a decision to participate was taken in 2004. A working party was formed, consisting of the Academic Services Manager, the E-strategy Co-ordinator, and Adrian Machiraju, now re-titled Information Consultant, who became the Project Officer.

During 2005 a repository was set up for Royal Holloway on the server hosted at UCL, and tests were carried out with a handful of EPrints volunteered by personal contacts of the working party members. The repository was officially launched as Royal Holloway Research Online (RHRO) in January 2006 at a buffet reception for researchers presided over by the Principal. Submissions were invited from any member of the institution, and in the same month a part-time student assistant was appointed to help with the record creation.

4.4.5.2 Eligible Content

The purpose of the repository is to increase access to Royal Holloway's research outputs, and the original intention was that any research, published or unpublished elsewhere, authored or co-authored by a member of Royal Holloway would be eligible for inclusion. This rule has had to be revised repeatedly to meet the needs of the community. Royal Holloway is an active member of a number of collaborative research programmes, and a strict insistence on authorship by members would have left the team unable to mount the complete outputs of these programmes. The current definition would be "published or unpublished research authored or co-authored by members of the College, carried out at the College, or as part of programmes organised by or in collaboration with members of the College".

The number of EPrints has now passed 500, and the great majority are journal papers, working papers, papers read at conferences, or chapters in books. All the material is text-based; as yet no audio-visual materials have been submitted, though the Department of Drama and Theatre Studies has expressed an interest in doing so.

4.4.5.3 Data Collection

To date it has been entirely voluntary to submit EPrints to RHRO. The Principal is keen to introduce a College policy that all research outputs should be deposited, but that was not in force during the period of this study. Researchers are encouraged to submit their draft papers as soon as they are finished, to ensure their preservation. A surprising number of authors do not preserve their drafts, and are then frustrated, as not many publishers permit the use of their edited e-journal files.

There are two means for authors to submit their work. A special email address for eprint submissions has been created, to which submissions can be sent as email inclusions, and the mailbox can be opened by any member of information services staff working on the repository. Royal Holloway authors may also create their own accounts and submit their papers directly to the repository, creating their own metadata. In the early days of open access, it was widely believed that self-deposition would greatly reduce the staff time required to run repositories. This does not prove to be the case. It is still necessary for repository staff to correct the metadata, as academic staff cannot be expected to be expert in ISSNs and DOIs, or in issues concerning copyright permissions. However, self deposition

does save some time, and gives researchers a much greater sense of involvement which is itself valuable.

For submissions which arrive by email, metadata records are created by repository staff, mostly but not always by the part-time assistant. All new submissions go into a buffer for quality checking before being made visible to the outside world. It was originally intended that the Project Officer would always do this checking, but the work of the part-time repository assistant proved so meticulous that she was authorised to do this also.

The copyright status of all submissions is checked by repository staff before they are made visible, even if the standard disclaimer has been clicked through by a submitting author. There is no doubt that the institution as publisher is liable for any breaches in English law. If no publisher's policy can be found, or the policy is not permissive, the author is contacted by email and asked if any rights had been specifically retained. If not, or where authors are uncertain, they are asked if they would mind requesting permission, using a standard form of words which is supplied. This procedure has been adopted because experience has shown that approaches from authors are much more likely to meet with a favourable response than approaches from librarians.

There have been several requests for the repository to scan in older papers from hard copies, so that staff can complete the collection of their work in the repository. However, the copyright position for making digital copies of print originals is particularly problematic, as there are issues with the quality of the scanners readily available and the scanning process itself is expensive in terms of staff time. As a result, this method has not yet been pursued.

4.4.5.4 Related Activities

There are three types of activity under this heading: advocacy, enquiries and maintenance.

Advocacy is the activity of explaining the concept and value of open access to productive members of College, encouraging them to make use of the repository. It is mainly conducted by the Project Officer, sometimes supplemented by other members of the Information Consultants' team. It is done by email, writing articles for internal newsletters and the intranet, attendance at faculty and departmental boards and other fora,and via publicity events which often include catering. It is therefore a relatively expensive activity, but the costs will drop markedly as the repository becomes an established part of academic life.

Enquiries come from both members of Royal Holloway, either as potential contributors or as users, and from users outside the institution. A dedicated email address is provided for them, and clearly displayed on the repository, partly so that any person wishing to object to anything published there can easily do so. No such objections have been received to date; outside enquiries have always been requests for more information.

Maintenance covers a miscellaneous range of operations on the repository itself. Examples would be updating EPrints with newer or corrected versions, either of the eprints themselves or of elements of the metadata; amendments to the subject tree as the institution's structure of departments and research centres changes; software updates and the installation of new software.

4.4.5.5 Formats

Submissions most often arrive as word-processed files, and are converted to PDF before mounting. This was decided as the PDF seems to be the most widely-accepted standard at present, readers are freely available and usually ready-installed on new machines in all platforms, and academic users are already familiar with the format from online journals.

Even though a free Word viewer is available from Microsoft, it was felt that mounting eprints in the format of a commercial word processor subject to frequent upgrades should be avoided. The PDF files are also secured from alteration, though of course it is known that this security is weak and easily circumvented.

As mentioned, no research outputs other than textual documents have yet been submitted for archiving, so the question of which formats to use for presentations, sound and videos has been deferred. The greater technical difficulty in converting these, and the extreme uncertainty about which formats are likely to be long-term standards, make it likely that such materials will simply be mounted in whatever formats are submitted.

4.4.5.6 Digital Preservation

For the period of this study, RHRO has been hosted on the server at UCL, so UCL has been responsible for the backing up of data. It is planned to transfer the repository to a server on campus in the summer of 2008, and a back-up standard for this has been agreed as part of the project specification. Longer-term preservation of digital objects is a serious issue, and RHRO is a member of the SHERPA DP2 Project which is examining this. Apart from supplying data to SHERPA DP2, there has been no effort in this area at Royal Holloway.

4.4.5.7 Lifecycle Costings

4.4.5.7.1 Background: Staffing

The only substantial and estimable costs in the creation of RHRO have been staffing costs, and even the staffing has been patchy and rather minimal, as the repository, being a new activity, has not featured strongly in the Library's sense of priorities in a period of marked staff shortages in its senior management. Of the three members of the original working party, the Academic Services Manager and the E-strategy co-ordinator were involved only in attending meetings and giving their views. No attempt therefore has been made to quantify the cost of their involvement, which may be regarded as part of their general managerial duties. It happens that both left Royal Holloway during the period of this study, and neither has been replaced, which has been one element of the staffing shortage mentioned.

The costs which can be estimated are those of the Project Officer and part-time repository assistant. The Project Officer, an information consultant, has spent an average of one tenth (10%) of his working time on the repository since its inception, rather more in vacations, very much less in the first term of the academic year. This cost has been estimated as 0.1 FTE of the full cost of employment in this post. The repository assistant has worked an average of six hours a week at £6.85 an hour, and the precise number of hours and amount paid was 285 hours and £1,925 respectively. The involvement of other staff, mainly information consultants, in answering questions about the repository from their liaison departments or referring them on, has been occasional and cannot sensibly be estimated.

4.4.5.7.2 Overall Approach

Costs for the year to December 2007 were totalled up and divided by the number of eprints added during the period. The division of time between the different aspects of adding an eprint was estimated by the Project Officer, sampling the last thirty eprints he added personally in 2008. This is because, at just the time this report was begun, the repository assistant was ending her employment, so could not be asked to sample her own working. The Project Officer has added some eprints personally throughout the repository's existence, and is content that the recent ones have not been in any way atypical.

4.4.5.8 Lifecycle Costs: Commentary

4.4.5.8.1 Creation or Purchase

Research outputs are created as part of the mission of the institution, and no part of the cost of creation is incurred by the repository.

4.4.5.8.2 Acquisition

The most relevant heading here is IPR & Licensing, which covers the costs of checking for publishers' copyright policies, and consulting the author and/or publisher if the policy is untraceable or ambiguous. Time for advocacy and publicising the repository may also be included in the "Negotiating of Submission" process, as without them the submissions would not occur. These activities are relatively expensive, as they are carried out by the Project Officer rather than the clerical assistant. During the year in question, two significant advocacy events for the whole of Royal Holloway were held, together with a number of visits to departments, and all the time spent in these has been included under "Negotiation of submission".

Selection, in terms of updating the collection, eligibility policies and the submission itself, did take some time, partly in discussing these issues at meetings and by email. However, the year studied was the third year of RHRO's existence. By this time these discussions had taken place, and the purpose of the repository in increasing access to research carried out at the institution, by or under the supervision of its members, was well-established. It may well be the case that for most repositories, policy definitions like these are non-recurrent startup costs.

IPR and Licensing policy and procedures, on the other hand, were ongoing, as the relevant law is unclear and mostly untested. The Project Officer was involved in discussions both with contributors and on the relevant professional lists, and attended a one-day workshop on the subject. This expenditure of time is reflected on the spreadsheet. The sending of enquiries about specific items, and recording of permissions, were carried out by the clerical assistant.

4.4.5.8.3 Ingest

The relevant heading under Ingest is Quality Assurance, for which the details are described above. This activity, including all checks made before eprints were mounted and their mounting, was largely carried out by the Project Officer.

4.4.5.8.4 Metadata Creation

Metadata creation is the second significant cost, with IPR and Licensing, applicable to every eprint added. It can include an amount of web searching to establish precise details of publication, about which submitting authors are sometimes vague or inaccurate. All this was normally the work of the clerical assistant.

4.4.5.8.5 Bit-stream Preservation

For the entire period of this study, RHRO has been hosted on the shared SHERPA-LEAP server at UCL, which has been maintained for the partnership by UCL, and supported by the SHERPA-LEAP Project Officer. There is no realistic way to divide up these costs between the SHERPA-LEAP partners, but it is estimated that the total cost per item is minute.

4.4.5.8.6 Content Preservation

As discussed in the section 4.4.5.6, Royal Holloway's only effort in this regard has been its involvement in SHERPA DP2.

4.4.5.8.7 Access

The only measurable heading here is for user support, which includes the checking of the EPrints and eprint-submission mailboxes. This activity was divided fairly evenly between the Project Officer and clerical assistant.

4. Institutional Repository Case Studies
SHERPA DP Case Study
SHERPA DP Mapping
SHERPA-LEAP Case Study
Goldsmiths
Royal Holloway
UCL

4.4.6 Case Study 3: UCL (University College London)

4.4.6.1 Background

UCL EPrints was founded in March 2004. The initial decision was taken to build the repository 'from the ground up', with the aim of building a critical mass of expertise, content and support, backed by usage data, in order to be able to demonstrate the value of the repository to the institution and to secure permanent recurrent funding. Seed funding came from SHERPA-LEAP.

4.4.6.2 Eligible Content

Any UCL-authored or co-authored research output is eligible for inclusion in the repository. Journal articles, conference papers and book chapters form a high proportion of the content. Working papers, theses, patents, reports and other outputs are also represented. Most of the content is textual, with only a few A/V objects currently held.

Supplementary data is also accepted; depending on size and audience, such data might be stored for curation in the Library's 'Digital Collections' repository (http://digital-collections.lib.ucl.ac.uk), with appropriate navigation in place between publication(s) and data. Digital Collections is a newly-introduced service, and does not, at the time of writing, store any data relating to UCL's EPrints collection.

4.4.6.3 Data Collection

Deposit with UCL EPrints is voluntary. Researchers are encouraged to deposit research outputs at the earliest stage of completion. For journal articles, this is the point of acceptance for publication, incorporating referees' comments: in effect, the author's final draft. The early capture of research outputs maximises their public lifespan, and so helps to increase their impact. Additionally, capturing research outputs at the pre-publication stage helps to overcome potential copyright barriers to local storage and dissemination.
Deposits are made by email, to keep the process as simple as possible for researchers. Papers are sent as attachments to a generic email address. Researchers are not permitted to upload metadata and objects directly to the repository: early experimentation found that this created too many quality control issues to be effective. A simple Web form, which will capture basic detail (e.g. depositing author's name and UCL Identifier, title of paper, publication, funder's code) in a structured way is in preparation.

On receipt of a research output, repository staff check the relevant copyright permissions, negotiating with publishers when appropriate, and create a full metadata record. Peer-reviewed records are distinguished from other deposits. Items 'in press' are marked as such; these are periodically reviewed and updated with full details (e.g. pagination) when available.

The digital object is, in most cases, converted to pdf before uploading takes place. Metadata-only records are accepted, typically where an author or department is keen that the repository should be able to supply a complete publications listing, or where copyright permissions prevent a full publication from being added to the repository. Around half of the records in the repository are currently metadata-only records.

Every new record is initially placed in a holding area for a quality assurance check by a senior member of staff. Typical checks include the copyright status of the research output in question, the bibliographic accuracy of the metadata, including any external links and identifiers, and the integrity of the digital object.

4.4.6.4 Related Activities

Much of the work of the repository staff involves outreach to the UCL community, both in publicising and marketing the repository, and in answering day-to-day questions from members of UCL. Repository staff also deal with enquiries from users of the repository, for instance about service availability and the reuse of deposited objects.

4.4.6.5 Formats

As noted under 6.1.3, conversion to pdf takes place in most cases (the few A/V files collected being the exception) as part of the content acquisition process. The rationale for this is two-fold: first, pdf files may be accessed using a free reader, and therefore the format is considered to be an appropriate one for an open access repository; second, it was felt that there was some merit in standardising to one format, both for consistency of presentation and to reduce some of the challenges of preservation in future. This policy is currently under review: it is hoped that the right balance between ubiquity, openness, 'mineability' and archival utility on the one hand, and the costs of file transformation(s) and implications for preservation on the other, will be found.

4.4.6.6 Digital Preservation

A Digital Preservation Policy for UCL Library Services is in preparation.
Meanwhile, adequate arrangements for bit-level preservation are in place. The Library has investigated outsourcing opportunities for the strategic preservation of the content of its EPrints repository (e.g. through participation in the SHERPA DP Project), and will continue to consider outsourcing solutions for the preservation of this material. Also under consideration is in-house preservation, making more use of the DigiTool platform which underpins the UCL Library Services Digital Collections service: DigiTool offers more sophisticated support for the long-term preservation of digital content than GNU EPrints, and will be used to support all the digital preservation activities which the Library undertakes in-house. However, it is felt that the formats so far acquired are well-known and stable (a recent collaboration with the PRESERV project confirmed that the content of the repository is not obsolescent), and that the Library has time to devise a thoroughly-researched and appropriate preservation strategy for the repository. Most of the resources available to the repository, therefore, are devoted to the more immediate challenges of embedding the repository in the day-to-day workflows of researchers and educating research authors about copyrights and open access. Preservation support will become both more pressing and more achievable when these challenges have been addressed successfully. This is not to advocate complacency, but as resources are limited, prioritisation has to take place.

4.4.6.7 Lifecycle Costings

4.4.6.7.1 Background: Staffing

The decision to begin the repository as a pilot and work towards strategic adoption by the institution has meant that the staffing quotient available to UCL EPrints has tended to lag behind the popularity of the service among UCL researchers. Various staffing permutations, involving portions of the time of permanent staff taken from other duties, supplemented by fixed-term, part-time data entry staffing using project monies, have been in place since the repository was inaugurated. Staffing has rarely been consistent from quarter to quarter since the repository's inception, but the overall trend has been towards longer-term staffing for the repository, in increased numbers.

The staffing costs used for this report are an approximation based on the costs of a full-time EPrints Assistant plus 0.2 FTE of a Manager. This is a reasonably close approximation of the staffing resources which UCL Library Services was able to utilise for the UCL EPrints repository during the twelve months preceding the preparation of this study.

2007 salaries, including on-costs:	
Assistant	£24,768
Manager	£55,759 x 0.2 FTE
Total repository staffing costs, year 1:	£35,920

4.4.6.7.2 Overall approach

The time spent by repository staff on the various elements of the LIFE model was determined during a period of sampling, which took place early in 2008, in as much detail as possible based on the information available at the time about the LIFE model. These findings were then retrospectively applied to the 2,266 records/objects which had been acquired over the preceding 12 months.

Non-storage costings were therefore derived from the following 3 inputs:

1. Number of objects ingested per year, using calendar year 2007 as an indicator
2. Estimated staffing expenditure on repository in the 2007 period
3. Findings from sampling to indicate how staff time is proportioned across various repository activities

Storage costings also necessarily involve some estimation:

1. It is assumed, for reporting purposes, that a similar number of repository objects will be ingested by UCL every year.
2. The annual cost of storage was estimated for these purposes at £1,116, based on estimated server/storage costs and an estimated 3-year refreshment cycle.
3. All costings shown in the lifecycle spreadsheet are per repository object.

The figures quoted should be regarded as estimates, and the Case Study itself as illustrative.

4.4.6.8 Lifecycle Costs: Commentary

4.4.6.8.1 Creation or Purchase

The costs of content creation are out of scope of this study. For these purposes, content is donated.

4.4.6.8.2 Acquisition

The main acquisition cost is in IPR & Licensing. There is also a modest 'obtaining' cost, associated with advocating the repository to potential depositors and dealing with their enquiries; and a check-in cost, in cases where acknowledgement of upload is made to a depositor.

While there are costs associated with selection - the occasional review of the repository's collecting policy and eligible content - and with maintaining the submission agreement, which is also subject to occasional review - those two costs are negligible when considered on a per-object basis. They are treated here as part of the general managerial component of all the staffing-based costings.

4.4.6.8.3 Ingest

The main ingest cost is in Quality Assurance. As outlined above, this takes place at a number of stages: a check against pre-existing records to eliminate duplication takes place before ingest begins; and a series of checks, both by depositing staff and by a second editorial reviewer, are made before the record is committed to the repository. The combined costs of file format transformation, for full-text objects, prior to deposit, are also incorporated here. Holdings update and deposit are automated; the cost of reference linking as defined by LIFE^2 is here absorbed into metadata creation.

4.4.6.8.4 Metadata Creation

Metadata creation is a substantial cost. The figure given here includes an initial search, sometimes across several sources, for external/authoritative publications data, which may or may not be discoverable and reusable. The figure also includes the periodic review of 'in press' items and their update with additional publication metadata, when available. Metadata QA takes place, but in this model has been treated as part of Ingest QA.

4.4.6.8.5 Bit-stream Preservation

Repository administration costs are here included in the general staffing costs.
Storage costs apply. It is not practical to detach the cost of the LEAP server/processor from the costs of its storage facility; server and storage are both small, and at per-eprint level, the actual storage costs would be miniscule. Storage costs for year one have therefore been calculated using one third of the server cost (based on an underlying assumption that the server is replaced every third year).

Note that the year one costs are shared across the total number of objects held in the centrally-hosted SHERPA-LEAP repositories. In subsequent years, the same server will hold only UCL objects. Two further assumptions have been made in calculating the storage costs for future years. First, as the growth rate of the UCL repository in the coming years is impossible to predict, it is assumed for Case Study purposes that a similar number of objects will be added to the repository each year. Second, neither the size of server/storage required in future, nor the costs of such hardware, can be predicted accurately, but it is assumed that the former will escalate while the latter shrink, and so, for Case Study purposes, a constant annual server/storage cost for UCL has been used.

Refreshment, as defined in the model, takes place periodically, but the costs are slight. Similarly, backup costs are in scope, but the process is automated, and the cost per item is negligible. Inspection does not currently take place.

4.4.6.8.6 Content Preservation

Preservation watch, as noted above (6.1.6), does not occur systematically. Preservation planning will, in future, include the periodic review of UCL's over-arching Digital Preservation Policy. Per repository object, this is a negligible cost. Preservation action, in so far as it is included in the Case Study, is attached to the costs of pre-emptive format migration at Ingest. The costs of preservation action thereafter cannot reliably be estimated; but given the stability of the formats currently in use, these costs could be nil within the 10-year scope of the model supplied for the Case Study. If an outsourced content preservation solution is implemented at UCL then that will provide a simple basis for per-object costings; for now, preservation is a small part of the repository expenditure.

4.4.6.8.7 Access

The main cost here is in user support, delivered through a telephone and email enquiries service, and through maintenance of the repository Web site. Access provision is automated, as is the logging and reporting of usage statistics. Access control, because of the way in

which it is implemented, is treated as part of Metadata Creation; but in fact it is so rarely applied as to be negligible for these purposes.

4.4.6.9 Non-lifecycle Processes and Costs

4.4.6.9.1 Management and Administration

Management costs within the scope of the Case Study (ie those specifically applying within the repository unit) have been treated as lifecycle costs and shared, in appropriate proportion, between the various components of the model. It is felt that these management costs are inseparable from the object lifecycle.

4.4.6.9.2 Systems/Infrastructure

As noted above, the systems costs which are within the scope of the Case Study (see section 3 for a definition) are so small, per item, that it is not helpful to treat them separately from the storage costs. Therefore, systems costs are here costed under storage.

4.4.6.9.3 Economic Adjustments (Inflation and Discount)

It is felt that inflation should feature as part of the lifecycle costings, wherever appropriate (for instance, wages and systems costs over time), rather than be treated separately. Discount may apply to one-off purchases, for example, server hardware, but beyond that, it is not likely to feature routinely in the core business of repository management, and has been disregarded.

4.4.6.10 Comments on the LIFE^2 Model

It was not possible for the contributors to provide much more than a snapshot for these Case Studies. The costs of running a repository over a fixed period can be roughly ascertained, albeit with some caveats because of the patchwork nature of the funding in most cases. Reliable forward projections are not possible in what is a new and fast-changing environment. In general, repositories are not yet securely embedded in their institutional research workflows: their roles, and the levels of institutional investment in their maintenance, are likely to change over time, with knock-on effects on costings.

It is clear, especially from the Goldsmiths experience, that different types of digital object within institutional repositories require different levels of attention and expertise in year one, and may require different levels of action and intervention in future years. The 'EPrints Repository' currently has broad scope, and a simple, per-object average across an IR may not tell the full story.

The QA of descriptive metadata is often a costable element of repository management, and, since Metadata Creation stands apart from Ingest in the LIFE model, an element of Metadata QA could perhaps be included under Metadata Creation.

The model helps to apportion the costs of preservation monitoring and action. The model is rightly neutral regarding preservation solutions (in-house, outsourced, migration, emulation, etc.). However, at this early stage, there seems little way to predict how IR managers might preserve their content over time: perhaps a national, outsourced, subscription-based solution will emerge? Might there be community-developed emulators or plug-in transformers for particular formats? How much local development work will be undertaken? It seems that, at present, only those potential users of the LIFE model who have in-depth experience of digital preservation of their particular types of content - in effect, those who already know what their

costs are - will be able to identify and allocate content preservation costings over time with any degree of realism. Perhaps some heuristics, even rudimentary, about the expected frequency of preservation action in relation to preservation strategy might be helpful for other users of the LIFE model: if such guidance is not already available then this may be a suggestion for future work.

The costs of post-ingest Appraisal, De-selection and Disposal are not included in the LIFE Model. The contributors felt that it may be helpful to consider acknowledging these lifecycle components in future iterations of the model.

The LIFE Model narrative explains that the model is designed to support comparability across different lifecycles and different institutions. The partners felt that such comparisons might be difficult to implement with any degree of reliability. Many costs, particularly those from year 2 onwards, are necessarily based on estimates. Even in this simple Case Study of institutional repositories, several potential differences between institutions emerged: costing methods differ between institutions; base costs differ - any staffing-driven process is likely to cost more in central London than in the rest of the UK, for instance; each repository can have a different remit, by which different types of object may be collected, with different levels of intensity; different levels of institutional support for repositories also have a bearing on operating costs. The LIFE model aims, commendably, to flatten out these differences, but the partners were concerned that it is insufficiently prescriptive to deliver a 'precise and repeatable' model in practice. Clearly, for instance, any direct comparison between London institutions and other UK institutions is not straightforward; how does comparison between, for example, UK and mainland European or non-European institutions stand up?

Regardless of their reliability, the partners were also unsure of the utility of inter-institutional comparisons based on cost per object. It is unclear in what scenarios such comparisons might be valuable. A higher-level, less detailed comparison, which might include staffing FTE, roles and grades, and repository size and remit, could in some circumstances be more insightful. The partners did agree, however, that the model may assist future service planning within an institution. It may support the preparation of costings to propose new services, or justify existing services within an institution. Comparisons between existing services within an institution may also be instructive.

In general, the partners suggest that some work on use cases (if not already undertaken) may be helpful in shaping the future development of the LIFE model and its applicability.

4.4.7 Key Conclusions

The costs indicated here should be regarded as illustrative, rather than absolute, because of several factors, among them the patchwork nature of repository funding; the need to base some costs on assumptions about future growth, expenditure, and preservation requirements; differences in costing methods and interpretations of the LIFE v1.1 model.

Table 32 - Overall Costs for SHERPA-LEAP Repositories

	Year 1	Year 5	Year 10
Goldsmiths	£31.50	£32.00	£32.20
Royal Holloway	£23.10	£23.60	£23.90
UCL	£15.00	£16.50	£16.70

The costs per object for Year 1 at Goldsmiths is £31.48, at Royal Holloway is £23.13, and at UCL is £15.98.

After 5 years, the estimated expenditure on each object is £16.45 at UCL, rising to £16.72 after 10 years, £23.60 at Royal Holloway, rising to £23.87 after 10 years, and £31.95 at Goldsmiths, rising to £32.22 after 10 years.

The variations in costings between the institutions may be attributed to three factors. First, the caveats already listed at 7.1 above apply. Second, the narratives show staff on different grades, in differing proportions, working in the repositories. This naturally affects the costings. As IRs become more stable, staff gradings and roles are likely to become regularised, and comparison across the HE community will become more informative. Finally, the studies show that the fact that Goldsmiths handles a widened range of digital materials within its institutional repository structure increases the average handling cost per object.

After year one, the main lifecycle costs are those associated with preservation. Bit-stream preservation costs are based on estimates, both of repository growth and the technology marketplace. Content preservation will clearly bring costs for the partners in future, but for the time being, those costs are not easily predictable.

The partners were unsure whether the model would, in practice, be able to meet the stated aspiration of providing a basis for inter-institutional comparison, and were equally unsure whether support for inter-institutional comparison should be the primary purpose of such a model.

4.4.8 Suggestions for Future Work

4.4.8.1 The LIFE Model

- The Project team might consider the inclusion of Appraisal, De-selection and Disposal as lifecycle elements.

4.4.8.2 Related Work

- Some guidance about the expected frequency of preservation action in relation to preservation strategy might be helpful for other users of the LIFE model.
- Some work by the Project team on use cases may be helpful in shaping the future development of the LIFE model and its applicability.

4.5 Section Review

This section has outlined some of the key issues surrounding the adoption of the LIFE Model from an IR perspective, as well as highlighting some of the overall lifecycle costs. A full breakdown of all lifecycle costs are given in the SHERPA DP and SHERPA-LEAP spreadsheets, and are summarised and discussed in Section 1.

5 BRITISH LIBRARY NEWSPAPERS CASE STUDY – COMPARING ANALOGUE TO DIGITAL COLLECTIONS

5.1 Purpose of this Section

This document outlines Work Package 4 of the Project. This work package examined the issues of digitisation as surrogacy through the British Library Newspapers Case Study.

The Case Study will follow the structure outlined here:

- Outline the background, terminology and aims
- Analyse the digital collection
- Analyse the analogue collection
- Describe the workflow
- Analyse the model's applicability to both formats
- Produce a comparison Table of lifecycle functions
- Produce a summary of lifecycle costs

5.2 Supporting Documents

There are several documents that support this Case Study, all of which are available from the LIFE Website:

- Two Excel spreadsheets are available with exact costings for the lifecycle of both the analogue and the digital collections used for this Case Study.
- Workflows for both analogue and digital collections were also developed. The Visio files for both of these diagrams are available for download, as well as being included as diagrams in this section (Figure 9 on page 80 and Figure 10 on page 83).

5.3 Background

The LIFE Model has changed to incorporate the latest thinking about digital preservation; this has an impact upon one of the project goals, the comparative evaluation of paper and digital collections. As with all Case Studies, this work package uses version 1.1 of the LIFE Model to compare an analogue collection of newspapers to a digitised collection of newspapers. It is expected by the LIFE team that being able to make such comparisons will help to inform future collection management decisions.

This work package will track the costs associated with the management of analogue and digitised newspapers at the British Library. It has used the Burney Digitisation Project as an example of a digitised collection of newspapers. This project gives a good overview of the type of lifecycle costs a digitisation project can expect to incur.

For comparative purposes the LIFE^2 Project has used the Legal Deposit of newspapers to provide the analogue costs. This is not a project, but a business function which was chosen due to its broad lifecycle activities. It was decided by the team that the analysis of this business process would bring the most benefit to the evaluation.

5.3.1 Definitions

The PREMIS2 definition[33] for an Intellectual Entity provides a useful starting point for the key terminology used in this report:

"Intellectual Entity: a set of content that is considered a single intellectual unit for purposes of management and description: for example, a particular book, map, photograph, or database. An Intellectual Entity can include other Intellectual Entities; for example, a Web site can include a Web page; a Web page can include an image. An Intellectual Entity may have one or more Digital representations."

Table 33 - Entity Descriptions

Intellectual Entity	Description
Issue	A complete issue of a particular newspaper
Page	A single page of a particular newspaper
Article	A single article of a particular newspaper

Within this report, an Intellectual Entity is considered to be a complete issue of a newspaper[34].

PREMIS2 also defines a lower level unit of a Representation. PREMIS focuses exclusively on the digital world, and has in fact chosen its terminology to avoid confusion with that used to describe analogue content. In the context of this project, it is useful to consider both analogue and digital content and workflows being described by consistent terminology.

"A Representation is the set of files, including structural metadata, needed for a complete and reasonable rendition of an Intellectual Entity."

The digital Representation of a Newspaper Intellectual Entity will closely follow the PREMIS2 definition, and might for example comprise a series of TIFF files for each page of the newspaper, and an XML file containing metadata. This will be termed the Digital Surrogate.

The analogue Representation of a Newspaper Intellectual Entity will comprise the original paper issue of the newspaper and associated microfilm[35] this will be termed the Analogue Object.

These definitions are summarised in Table 34.

[33] PREMIS2 Data Dictionary, pages 6 and 7, http://www.loc.gov/standards/premis/v2/premis-2-0.pdf
[34] It should be noted that this concept is explored further with regard to consideration of lower level entities, such as the article level or page level, later in the report.
[35] Details of the microfilm created at the BL, and a justification for its inclusion in the logical grouping of the Analogue Object are provided later in the document.

Table 34 - Definition of Terms

Entity	Representation	Description
Intellectual Entity	Digital Surrogate	The digital files comprising a specific issue of the Newspaper Intellectual Entity
	Analogue Object	The paper copy and associated microfilm comprising a specific issue of the Newspaper Intellectual Entity

5.3.2 Aim

The primary aim of this work is to evaluate whether the LIFE Model can be used to capture the costs of both analogue and digital lifecycles using examples from the chosen collections. If this can be achieved then the secondary aim is to compare the costs of both lifecycles at the same stages of the LIFE Model to analyse where the costs are similar and different.

5.4 The Burney Collection

The Burney Collection is a collection of Newspapers purchased from the Reverend Dr Charles Burney in 1818 for £18,500 with some additions made in subsequent years by the British Museum. It comprises over 1,100 volumes (190,256 issues) of the earliest known newspapers in the history of printing. These 1,100 volumes in turn comprise close to 1,000,000 pages of text from the 17th and 18th centuries.

Due to its age and its rarity, the collection has been managed through its analogue lifecycle by The British Library's curatorial and collection care staff. At various points in the collection's history decisions have been taken to extend the collection's life and to widen access for research and other uses. The two main decisions that LIFE² is interested in are the decisions both to microfilm and to digitise the collection. Both of these actions will form part of the digital lifecycle for this Case Study. It is important to keep clear that even though Burney is an analogue collection of Newspapers, it is the digitised Burney content that will be used for comparison to the analogue Legal Deposit of newspapers for the purpose of costing and analysis

Due to the Burney Collection's age, the original Newspapers are in a condition that means that re-scanning or re-microfilming is high risk. This is likely to become a common scenario for libraries as analogue collections age and means that the long-term management of the surrogates takes on a more important role within the lifecycle. The surrogates can in fact extend the life of the original object by limiting its use. This critical link between the two objects' lifecycle (analogue and surrogate) and the fact that the surrogates become the primary access objects means that the two objects are so closely tied that the team considers the surrogates as being part of the lifecycle of the object.

The microfilm for the Burney Collection was filmed in the 1970s and the digitisation was started in 1995-96 and ran until 2004. From 2004 onwards a project began to add value to the digital files by enhancing the metadata. It is this project, from 2005 to 2007, that is used within the Case Study to add cost to the digital versions.

5.4.1 Burney Digital Collection

The analysis of the digital surrogates has shown a number of common characteristics found in early digitisation projects. These characteristics which can sometimes complicate preservation action can be broadly placed into three categories.

1. older scanned versions of file formats (e.g. bi-tonal)
2. non-standard metadata
3. legacy storage media

5.4.1.1 The Master Source Files

There are 920,335 uncompressed bi-tonal TIFF image files, loaded in a directory structure across two logical drive partitions ('Burney' and 'Burney1') which are stored on a local server within The British Library ('W2k3-nasburney1').

These files take up 2,128 gigabytes of storage space and were originally derived from 114 DLT tapes. Generally each TIFF represents one scanned newspaper page, although there are some exceptions. The files were backed up onto LTO3 tape, as well as being stored on the project team's server.

5.4.1.2 The Service/Production Files

The source images have been cropped and de-skewed, then compressed via CCITT 4. There are both page-level images and article-level images, plus corresponding XML files which contain the OCR output and metadata. There is a one-to-one match (XML to TIFF) in the number of files at page level, but at article level the number of TIFF images is higher because a single article (one XML file) may span several pages (several images).

The service TIFF and XML files are also laid out in a directory structure, but this is not the same structure as the master source files. However each XML file contains a data tag with the file name and directory path of the TIFF file from which it was derived – this is the link back to master source.

The file numbers and sizes of the service output are reported in Table 35.

Table 35 - Burney Digital Files

	Number	Size
TIFF page files	916,652	263 gigabytes
TIFF article files	1,878,234	258 gigabytes
XML page files	916,652	140 gigabytes
XML article files	1,534,068	156 gigabytes

The XML data conforms to a project DTD, which is consistent across both the Burney and 19th century newspaper collections (different tags are used for each, but both are defined within the same DTD).

All of this data is also on LTO3 tape (eight tapes) sent from the contracted digitisation supplier. There is a third copy of the data on a separate (W2k3-nasburney1) server as well.

The digitised content is typical of digitised projects from this period. The image content is captured in TIFF files with supporting information in associated XML files. A digital

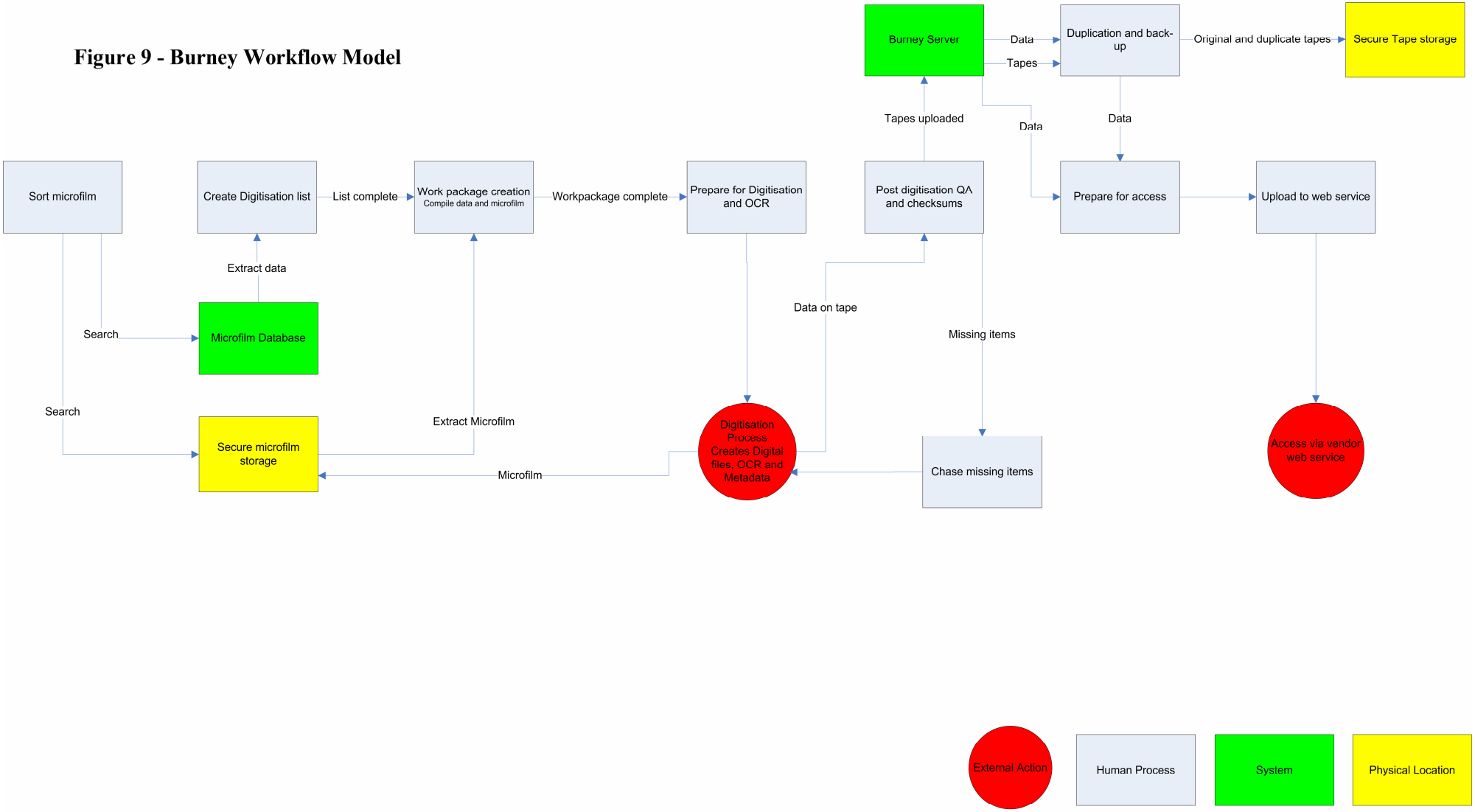
preservation technology watch activity is already in place at The British Library for TIFF and XML. Both are considered low risk at present due to the widespread industry support for both formats. It is also worth mentioning that The British Library's digital preservation team are already working towards a solution for this type of content through its risk assessment and content stabilisation work.

5.4.2 Burney Workflow Model

Interviews with the Burney Digitisation Project Manager have indicated that the workflow for the project is as illustrated below. The costs for each process and function will be estimated based upon the Manager's experience of the time taken for each stage of production. These costs will be used to give examples in later sections and to populate the relevant LIFE v1.1 Model stage and element fields.

The workflow follows a fairly conventional route for digital projects. There is some organisation of the material into batches for production and from there a series of procedures is carried out by the team involving sending the items offsite to the contractor and tracking the digital files as they return to the project manager. There is also a backup procedure that is managed by the project team once the digital objects are securely in the repository.

Figure 9 - Burney Workflow Model



5.5 Legal Deposit of Newspapers

The British Library receives a copy of every national newspaper daily. As well as this it also receives the majority of regional daily and weekly newspapers. It receives these newspapers under Legal Deposit legislation and the material is managed by staff located at Colindale, North London. The team is responsible for handling 133,000 issues per year and the costs for the analogue part of this Case Study will use one year's figures from this operation.

The staff fulfil a number of functions in relation to managing this collection. These functions include collection management decisions for acquisition, storage and preservation as well as operational functions such as the movement of the objects within the store and administrative duties.

The workflow is quite different from that for the Burney digitised objects and this is reflected in the workflow diagram in this section. It involves more processes and the movement of the analogue objects (both microfilm and newspapers) between the two buildings based at Colindale.

5.5.1 Building 120

Building 120 is the main focus for LIFE^2 in this Case Study as it contains the majority of the activity relating to the Legal Deposit of newspapers. Building 120 is the deposit point for all UK newspapers sent to The British Library under the terms of the Legal Deposit Act[36] .

Building 120 handles 300 new newspaper titles and 133,000 individual newspaper issues every year. The building's purpose is twofold:

1. To safely store and manage Legal Deposit newspaper material
2. To provide some additional storage for other content on demand

5.5.2 Colindale Newspaper Library

This building is the better known of the two facilities run by The British Library at Colindale. It is the home of the Newspaper reading room where access is provided to the Library's newspaper collections for research or the general public. It is also the home of the Library's imaging team who undertake preservation microfilming work for newspapers. This activity is an important part of the analogue lifecycle and was studied in order to obtain a number of key lifecycle costs.

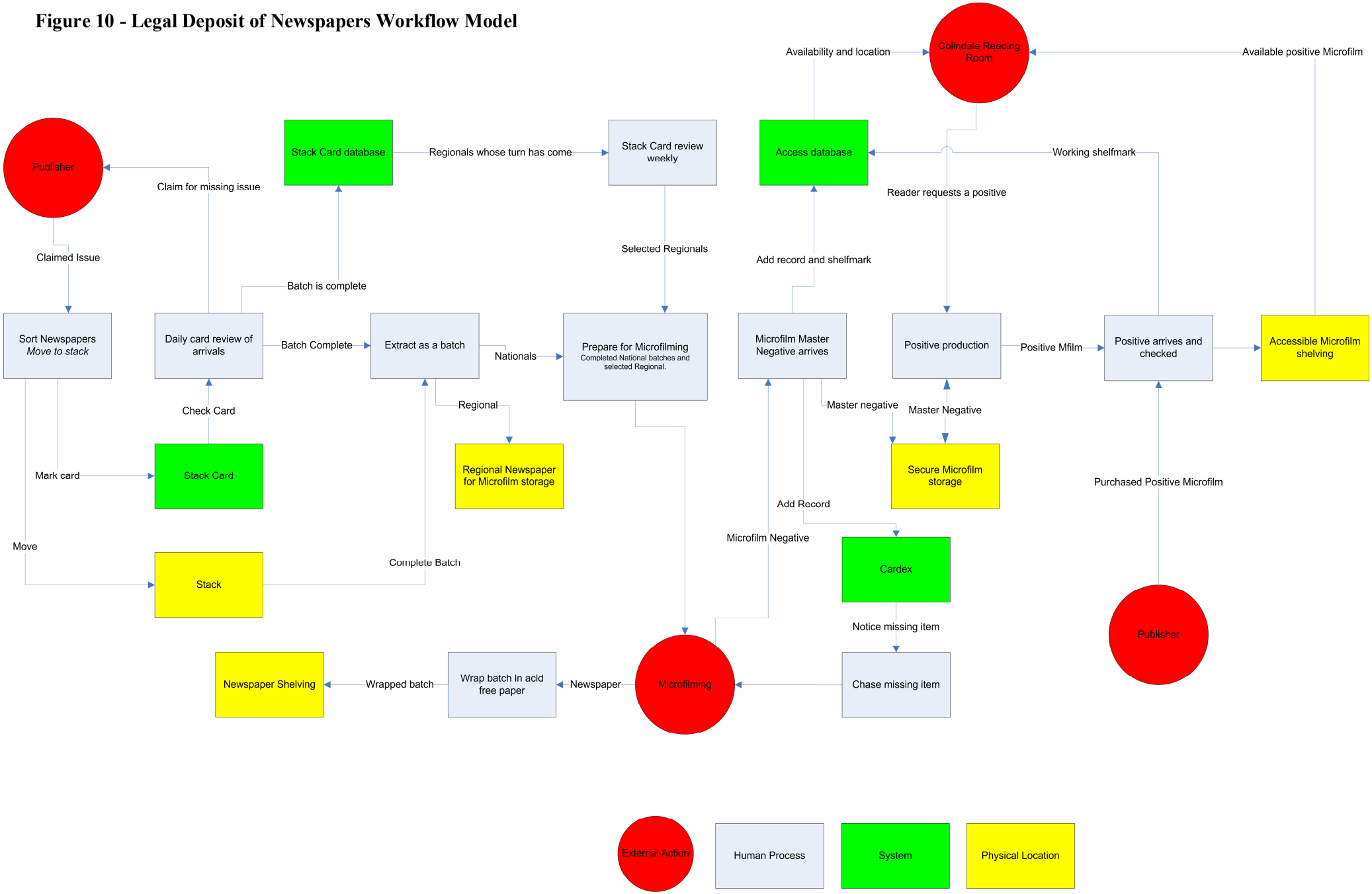
5.5.3 Legal Deposit of Newspapers Workflow Model

The workflow for the newspaper stream is a more complicated model than that of the Burney Digitisation Project. Interviews with the Manager at building 120 and other library staff have shown that the workflow is a mixture of manual procedures and some more automated systems. Procedural and system costs from this model will be used to give examples in section 4 and to populate the relevant LIFE v1.1 stage and element fields to produce final costs.

[36] Further information is available here: http://www.bl.uk/aboutus/stratpolprog/legaldep/

The workflow (Figure 10) starts with delivery of analogue newspapers to the loading bay of building 120. From there it follows a series of procedures outlined in the diagram below which involve the creation of microfilm surrogates and the reshelving of the analogue object into long-term storage.

Figure 10 - Legal Deposit of Newspapers Workflow Model



5.6 Using the LIFE Model v1.1 for Comparison

The LIFE Project is primarily concerned with the cost of the digital lifecycle and the first LIFE Project went some way to proving the applicability of the model to a range of digital lifecycles[37]. However comparisons between both analogue and digital lifecycles are crucial to making future collection management decisions. For example, when faced with the decision to acquire an analogue or digital version of the same object which one provides the best solution in terms of cost and sustainability? To help identify solutions to these questions, we used the LIFE Model to provide:

1. A direct cost comparison between paper and digital formats
2. A possible method for supporting decision making so libraries may decide for themselves what to keep when space or cost is a concern

5.6.1 Interviews with Key Personnel

The first challenge was to ascertain whether the terminology used for a predominantly digital preservation project could be used for an analogue collection. To achieve this, a group of British Library preservation experts were interviewed to gauge reaction to the terms used within the model and also to suggest any changes that might need to be made. The people involved in this process were:

- Stephen Morgan - British Library Collection Storage
- Deborah Novotny - Head of Preservation
- Dawn Olney - Head of Collection Storage
- Richard Davies - LIFE[2] Project Manager
- Rui Miao - LIFE[2] Project Assistant

A number of subject and functional specialists were also interviewed, and they included:

- Bhavna Tailor - Manager Legal Deposit of Newspapers
- Richard Gibby - Ex-Project Manager for Burney (now Legal Deposit Group Lead)
- Lucy Evans - Serials Acquisition Manager
- Ed King - Head of the Newspaper Collection

Discussions for this analysis were in four parts:

1. Is the LIFE Model terminology appropriate to use when identifying the costs associated with analogue collections?
2. Is the stage level definition applicable to analogue collections?
3. Is the element level definition applicable and understood when working with the costs associated with analogue collections?
4. Is the sub-element definition applicable and understood when working with the costs associated with analogue collections?

The results of these interviews were collated and are summarised an element at a time in the sub-sections below.

[37] Further details on the Case Studies are available from the LIFE[1] Project Report - McLeod, R., Wheatley, P. and Ayris, P. (2006) *Lifecycle information for e-literature: full report from the LIFE project.* Research report. LIFE Project, London, UK. Available online: http://www.life.ac.uk/1/documentation.shtml

5.6.2 LIFE Model v1.1

Figure 11 - LIFE Model v1.1

Lifecycle Stage	Creation or Purchase[1]	Acquisition	Ingest	Metadata Creation[2]	Bit-stream Preservation	Content Preservation	Access
Lifecycle Elements	. . .	Selection	Quality Assurance	Re-use Existing Metadata	Repository Administration	Preservation Watch	Access Provision
	. . .	Submission Agreement	Deposit	Metadata Creation	Storage Provision	Preservation Planning	Access Control
	. . .	IPR & Licensing	Holdings Update	Metadata Extraction	Refreshment	Preservation Action	User Support
	. . .	Ordering & Invoicing	Reference Linking		Backup	Re-ingest	
		Obtaining			Inspection		
		Check-in					

The stage and element definitions for the model in Figure 6 are given in section x.

5.6.2.1 Creation and Purchase

The Lifecycle Stage of Creation or Purchase needs no change.

5.6.2.2 Acquisition and Ingest

The Lifecycle Stages of Acquisition and Ingest need no change.

5.6.2.3 Metadata

No changes recommended.

5.6.2.4 Bit-stream Preservation

Bit-stream preservation is a new distinction within the LIFE Project. It relates to the costs associated with the bit-stream calculated separately from the content. It is a specific term used within the digital preservation community and has little meaning within the analogue world. After discussion with the LIFE team the term 'book storage provision' was used. It was useful that the acronym (BP) was the same which simplified the identification of cost for the analogue object detailed below and allowed the team to cross check with the digital surrogate.

5.6.2.5 Repository Administration

Administration functions of the Legal Deposit newspaper collection are commonplace. There are administration systems within all analogue and digital lifecycles.

5.6.2.6 Storage Provision

No changes recommended.

5.6.2.6.1 Refreshment

This becomes the cost of refreshing the storage area. The group felt that it would not be a meaningful comparison if all costs were added (such as painting, damp proofing etc.) and that the comparison here with the analogue world should only be the reshelving of the newspapers after preservation work (microfilming) has been carried out. This was thought to be a closer comparison to the digital objects, where the cost is the movement of data from one location to another as part of a hardware refreshment activity.

5.6.2.6.2 Backup

The backup for Newspapers is microfilm and more specifically the master negative. The digital comparison for Burney is the tape backup produced for the project.

Note: At this point, it is worth documenting that microfilm preservation is undertaken in three stages:

1. Master Negative - The main preservation copy of the newspaper. Safely stored offsite from all BL buildings.
2. Duplicate Negative - Backup copy sent to the Newspaper Library at Colindale for storage
3. Positive copy - Primary access copy for use in the reading rooms

As mentioned previously in this Case Study, the LIFE team has reached a decision that microfilm is part of the analogue lifecycle rather than separate from it. This is due to the link between the analogue object's lifetime and the use of the surrogate as a method of extending access to it. Surrogacy in digital objects might comprise the access copies created in order to keep tight control over master files. In this way microfilm is considered to be a working copy of the original in the same way that a JPEG file might represent a Master TIFF in a digital lifecycle.

5.6.2.6.3 Inspection

Inspection is the process of ensuring stored objects can be retrieved without loss. Inspection is a broad-enough term for both physical and digital objects. A combination of manual retrieval and viewing will be used here.

5.6.2.7 Content Preservation vs. Conservation Procedures

This stage includes the cost of future planning activities and the predicted cost of keeping an object accessible. For analogue objects, this remains the same and it was decided by the group that Conservation Procedures was a meaningful term.

It is the conservation department's role within a library to conduct a variety of procedures all relating to keeping an object in a sufficiently-conserved state so that access can be maintained. These techniques include repair, maintenance and auditing of the object's properties. Again this also meant that the team were able to retain the acronym (CP) for this section which made comparison of costs more easily identifiable for this section.

The group made the comment that the LIFE term 'content preservation' was a very close match to 'conservation procedures' as a descriptive term. The term conservation procedures seemed adequately to describe the costs incurred by the conservation and preservation teams' role to plan for future action.

5.6.2.7.1 Preservation Watch

Preservation Watch remains unchanged and becomes a direct cost for the analysis of new paper conservation treatments and looking at new techniques for the future preservation of newspapers. This would comprise a comparison of the time used for Digital Preservation activities such as Technology Watch or monitoring the community vs. Preservation surrogacy watch activities such as monitoring the Cellulose Acetate Microfilm Forum (CAMF)[38] or the Image Permanence Institute[39] forums.

5.6.2.7.2 Preservation Planning

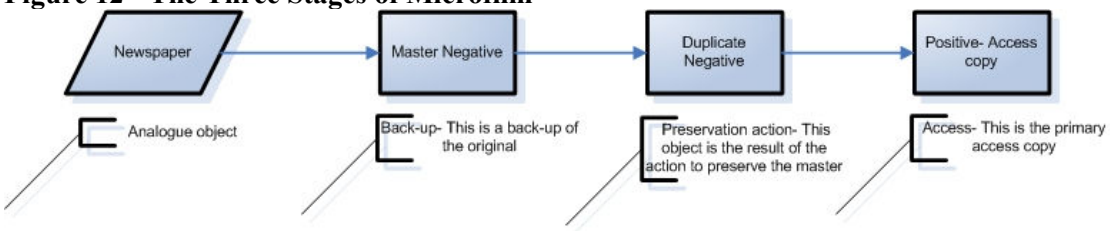
The British Library's preservation department run preservation planning activities for the Legal Deposit facility at Colindale. An example is the planning time taken to decide what to microfilm for the next 12-month period.

5.6.2.7.3 Preservation Action

For digital materials there are many different actions that may be taken - for example Migration, Emulation or Migration on demand. There is a strong comparison here with preservation and microfilming. As previously indicated, microfilming comprises three phases: Master Negative, Duplicate Negative and Positive. All three are separate functions. Where the Master Negative was considered to be a Backup of the original Newspaper, the Duplicate Negative (also analogue) is considered the result of the preservation action of the Master.

The comparison becomes the time taken for analysis and movement of the digital object with the microfilm duplicate negative which is the result of the action of preserving the original master negative. A diagram of the three stages and the role each plays in the lifecycle is given below (Figure 12).

Figure 12 - The Three Stages of Microfilm



5.6.2.7.4 Re-Ingest

This sub-element will form the final stage of the process where the checking and validating of the files from the Burney Collection will be compared with the checking and shelving of the duplicate negative microfilm into the Newspaper Library.

5.6.2.8 Access

There are strong crossover points between analogue and digital objects at the high level in terms of the components of access. It is expected that the definition used will require no changes.

[38] http://www.bl.uk/services/npo/journal/3/camf.html
[39] IPI is a non-profit research laboratory devoted to the preservation of visual material, see http://www.imagepermanenceinstitute.org/

The example for comparison is access to the digital object provided through software and hardware, whereas for the analogue object it is the positive master microfilm (the third and final stage) and the viewers within the Newspaper Library.

5.6.3 Conclusions of Evaluation

The analogue expert group were positive about both the terminology and the application of the LIFE model to analogue collections. General consensus on the four main questions posed can be summed up as follows:

1. Q. Is the LIFE Model terminology appropriate to use when identifying the costs associated with analogue collections?
   A. Yes. Interviewees felt that the definitions and guidance in v1.1 of the model could be followed.

2. Q. Is the Stage Level definition applicable to analogue collections?
   A. Yes. With the name change to the stages 'Bit-stream Preservation' and 'Content Preservation' to 'Book Storage Provision' and 'Conservation Procedures'.

3. Q. Is the Element Level definition applicable and understood when working with the costs associated with analogue collections?
   A. It is certainly workable for Newspapers. Other collections may need to consider changes to element definitions. It is recommended that more types of physical objects are assessed.

4. Q. Is the Sub-element definition applicable and understood when working with the costs associated with analogue collections?
   A. No. Specific library operational terms would be used at this level

The conclusion from this analysis is that the LIFE model v1.1can be used to describe both analogue and digital lifecycles in a meaningful way to both analogue and digital experts. At a high level there is strong confidence in this approach, but this confidence level drops as the level of detail examined is increased. Where the high level LIFE Stages are considered, there is a strong mapping between the analogue and digital processes. There is also a good correlation between analogue and digital at the Element level, within the Newspaper Case Study. Beyond this content area, further study and comparison may be required. At the Sub-element level there is little direct correlation. Given that the Sub-elements are provided as guidance only for digital lifecycles, it is unsurprising that they were found not to be useful for describing analogue lifecycle functions.

5.7 Comparison

In this section lifecycle functions will be used from the work flow diagrams in section 2 and 3. These lifecycle functions will be allocated a description based upon activities identified by the respective departments or projects. The descriptions have been established by the LIFE team using the appropriate information provided by the British Library's administrative departments. The lifecycle functions are measured in such things as time, people, purchase or unit cost of production and are placed side by side to show how the functions of costs of analogue objects and digital objects compare.

5.7.1 Table of Lifecycle Functions to Compare

The breakdown of activities given here is an example of how we have used the LIFE Model v1.1 to analyse the costs specific to both collections. This type of approach was found to be useful when used in conjunction with the workflow diagrams (Figure 9 (Burney Digital) on page 80, and Figure 10 (Legal Deposit) on page 83). The team has used the workflow diagrams to identify the lifecycle functions and has used this Table to place descriptions of the functions side by side for comparison.

Table 36 - Comparison of Lifecycle Functions

Lifecycle Stage and element	Legal Deposit of Newspapers	Analogue Costs	Burney Digital	Digital Costs
Creation or Purchase				
Creation	NA	NA	Cost to digitise the total archive plus create associated project information	Total project costs to digitise
Purchase	NA	NA	Purchase of Newspaper collection	Purchase price in 1818. However in today's market the materials would come to the British Library under Legal Deposit.
Donation	NA	NA	NA	NA
Acquisition				
Selection	Check the claimed issue arriving from publisher	Staff time by percentage	Sort microfilm into correct order and batches for scanning	Staff Time by percentage
Submission Agreement	Marking of the analogue card system as per agreement between Manager and staff	Staff time by percentage	Drawing up of contract with scanning contractor	Legal department time by percentage and Project Manager time
IPR & Licensing	NA	NA	Management of contract agreement with scanning contractor	IPR staff time by percentage
Ordering & Invoicing	Claim for any missing issues that were expected as part of the deposit of newspapers	Staff time by percentage	Compilation of the list to order digital files based upon the batches of microfilm selected	Staff time by percentage
Obtaining	Movement of newspapers to the book stack	Staff time by percentage	Create work package including compiled list plus microfilm and send order	Staff time by percentage
Check-in	Checking procedures to ensure that all newspapers are present. If	Staff time by percentage	Receive work package from scanning supplier and do quick check to confirm safe return of microfilm	Staff time by percentage

	newspaper not present this initiates a return to ordering and invoicing			
Ingest				
Quality Assurance	Review all arrivals to ensure Newspapers are complete and not damaged	Staff time by percentage	Detailed check of work package digitised scans including fixity values and virus checking	Staff time by percentage
Deposit	Batch marked as complete on stack card and recorded as having been deposited with date	Staff time by percentage	Time taken to move received scans and meta files from LTO tapes to server and to run final check to make sure the digitised content has moved by file count	Staff time by percentage
Holdings Update	Update computer database to add deposit record	Staff time by percentage	Update server database to add record of deposit	Staff time by percentage
Reference Linking	Serials Acquisitions team estimated time to create a catalogue record per newspaper issue	Staff time by percentage	NA	NA
Metadata				
Re-use existing Metadata	Meeting with staff to review stack card information and transfer information between batches of deposit	Staff time by percentage	Estimated time to export existing metadata from holdings update to Library system (Aleph)	Staff time by percentage
Metadata Creation	Request form for Microfilm order is compiled.	Staff time by percentage	Additional information added by project team to BL holdings records. Arrival date, signoff as being complete, date of QA check and name of person checking	Staff time by percentage
Metadata Extraction	NA	NA	Checking of extracted OCR for accuracy. Also verification of extracted metadata prior to access via scanning contractor website.	Staff time by percentage
Bit-Stream Preservation[40]				
Repository Administration	Time spent to separate the collection into National and Regional workflows	Manager time by percentage	Manager's time checking Burney server to confirm all expected content is present. This is a check that is done as an administrative function rather than as a detailed QA or integrity check	Manager time by percentage
Storage	NA (Estates	Estates cost	Burney project server hard drive	Hardware cost

[40] Maps to Book storage provision for analogue

Provision	costs do exist; however the per-linear-metre cost includes building infrastructure costs)		purchase costs	
Refreshment	Newspaper batch wrapped together and placed back into storage after microfilming.	Staff time by percentage	Time taken to manage the server and the content throughout the project life. Tasks include adding new hardware, checking hardware (checksum) and admin functions provided by IT support team	Staff time by percentage
Backup	Creation of master microfilm negative from newspaper	Unit cost from Imaging services	Cost of creating LTO3 tapes for backup	Staff time by percentage and some hardware
Inspection	Master Negative arrives and checked	Unit cost from Imaging	Visual inspection of digitised files to ensure that they can be opened and viewed. This check is done as a percentage of the total files received	Staff time by percentage
Content Preservation[41]				
Preservation Watch	Time allocated to preservation watch activity for paper and microfilm, involves checking of industry websites for most recent information	Staff time by percentage	Time allocated to preservation watch activity for the file formats used in the Burney project by The British Library's Digital Preservation team	Staff time by percentage
Preservation Planning	Preservation planning by the Preservation Surrogacy Masters Group which plans for future preservation activity such as microfilm	Staff time by percentage	Time taken to create preservation plan for the Burney project based upon the preservation watch mechanism and also using analysis to inform the plan provided by the technical lead highlighting any concerns or issues	Staff time by percentage
Preservation Action	Creation of the Duplicate microfilm negative. An action which creates a surrogate of the master, thereby protecting the original from overuse	Unit cost from Imaging	Time estimated to be taken by the Digital Preservation Team to secure the archive by running analysis of existing content, going back to original tapes to address missing/corrupt content and potentially moving the project information to the digital preservation store	Staff time by percentage
Re-ingest	Ingest the	Staff time by	Time estimated by the Digital	Staff time by

[41] Maps to Conservation Procedure for analogue

	checked Duplicate microfilm into the book stacks as the working copy	percentage	preservation team to gather new files from preservation action, compile report confirming completion of the preservation action and re-ingest into either the Intermediate Store or the Digital Library System	percentage
Access				
Access Provision	Creation of Positive Microfilm which becomes the primary access copy available in the reading rooms	Staff time by percentage	The cost of negotiating the specifications of the web-site with the scanning contractor and setting up the web applications to view the digitised files	Staff time by percentage
Access Control	Management of the database which controls the availability of the positive microfilm	Staff time by percentage	The cost to manage the project's digital rights based upon the licensing agreements agreed in the project documentation. This involves discussions with other potential partners.	Staff time by percentage
User Support	Reading room, reference staff to support access to the positive microfilm, advice and guidance to the general public and researchers	Staff time by percentage	The cost to support users who require access to the digital objects. This has involved training of key staff to use and show the collection. It also involves some administration of access to the content as there are limitations to its use	Staff time by percentage

5.8 Discussions and Decisions

In this section the LIFE Project team have captured the decisions made which have allowed them to get to the point of identifying cost-bearing functions which are described in the Table above. From here, the team has proceeded to put a final cost to these lifecycle functions to enable them to draw the conclusions, which are set out at the end of the document. Decisions which have had an impact on the final costs are captured here.

5.8.1 Growing and Static Collections

The Legal Deposit of newspapers building receives 133,000 issues every year and is an operation which is constantly ingesting analogue objects. The Burney digital project by comparison is not growing and has in fact ended. So one collection is getting larger and more complicated while the other remains the same size. For this reason it was decided to take a snapshot of the analogue objects after year 3 and to freeze the size of the collection. The decision to use years 1-3 for costing purposes was based upon the fact that the Burney digitisation project ran for three years.

This does not mean that all costs were frozen for the subsequent years (4-10) as many of the costs would still be incurred, particularly in terms of collection management and digital preservation, so these are recorded in the spreadsheet.

These decisions have helped make the comparison of analogue to digital costs more meaningful (like for like), but it is recommended that future comparison studies are undertaken using collections that are either both growing in size or where both have reached their natural size. This does not mean that the comparisons undertaken for LIFE^2 have been in vain, as the challenge for this research was to be able to use the LIFE model to identify the costs associated with an analogue collection and a digital collection. This has been achieved. The next stage in the comparison of the two collections is where the recommendation concerning growing or static collections becomes more relevant. The LIFE^2 team believe that, since the work is now done to identify the costs, a closer comparison of analogue and digital lifecycles is possible.

5.8.2 Creation or Purchase

For Legal Deposit newspapers there is no cost under this head due to The British Library's legal responsibility to take receipt of all UK Newspapers.

For the digital Burney collections, there was an initial cost in 1818 to purchase the original newspapers. The LIFE team, in consultation with our economic advisor, decided to withdraw this cost based upon the fact that if acquired today this collection would fall under Legal Deposit legislation incurring no cost.

For the digital Burney collections there was a substantial cost to digitise or create the data. The costs for this are based upon scanning and OCR costs stemming from the technology that was available at the time. In today's monetary terms the cost per page has dropped significantly and it was felt that a more up-to-date project cost should be used alongside the actual Burney Creation cost.

For this reason the JISC-funded "JISC1" British Newspapers Project has been consulted. This project has been digitizing 19th Century newspapers for the last two years and the scanning costs for this project will be added to a separate table in order provide an indication of change of digitisation creation costs over time. Statistics from the JISC1 Project will also be used to provide the average number of pages per newspaper for the analogue object comparison.

5.8.3 Entity, Object and Page Level

As stated in the introduction to this Case Study, an entity is considered to be any analogue object (i.e. newspaper or book) or digital surrogate.This definition has enabled us to identify the costs associated with each stage of the lifecycle for two different types of object. However, it is believed that as well as identifying costs per entity, a per-page cost would be a useful comparison, particularly when trying to single out the cost of creation for digitised archives.

For the Burney digital collections this was quite easy, as the team knows there is digitised content for 916,652 pages in TIFF and XML files. Each digital object comprises a TIFF file and an XML file.

For analogue materials, the arithmetic is much harder. Conversations with Collection area staff and Operational Microfilming staff led the team to believe that an average number of pages over different titles is misleading.

This is certainly true of 20th and 21st century newspapers, which vary in page length from day to day or week to week for a particular title. Newspapers from the 18th and 19th centuries did

not vary too much in length, and in fact many Burney newspapers started as one-page productions before increasing to two, four or eight pages. This is a far cry from the 100-page Sunday editions regularly seen today. So the challenge for page level information was how to establish a meaningful comparison. The largest project with the most information on digitised newspapers is the JISC1 Project. This project has had much experience in microfilming and digitizing newspapers from the 19th century. The project did not capture issue numbers, but it did measure page numbers. The project microfilmed and digitised volumes of material and each volume comprises a number of issues bound together. This means the LIFE team has page- and volume-level information, but no issue numbers. Estimation of the number of pages per issue is also difficult to establish. The most important point is that the team has some comparison at page level to help to isolate costs for creation, so that it can be seen whether creation costs are increasing or decreasing in the future.

Table 37 - Summary of page level information

	Issues	Total pages	Average pages per issue
JISC1 Newspaper Project	Not measured	2,051,127	Not measured
Digital Newspapers (Burney digital collections)	190,256	916,652	4.8*

* This figure of 4.8 pages to an object is used as the divisible for total unit cost.

Table 38 - Description of Comparisons to establish object and page level information

			Comparison	Aim
Object level	Analogue Newspaper object	Burney digital object	Total object lifecycle costs analogue vs. digital	Per object cost
Page level	JISC1 19th century newspapers	Burney digital object	Per-page cost digital vs. digital	Per page cost for digital.

5.9 Costs

Now that the discussions and decisions have been captured within the Case Study, a summary of all project costs can be produced in this section. The discussion starts with the comparison of total project costs for analogue and digital materials by Lifecycle Stage. Following this are costs at the entity, object, page and article level.

5.9.1 Total Project Costs

The costs within this section come from the spreadsheets that accompany

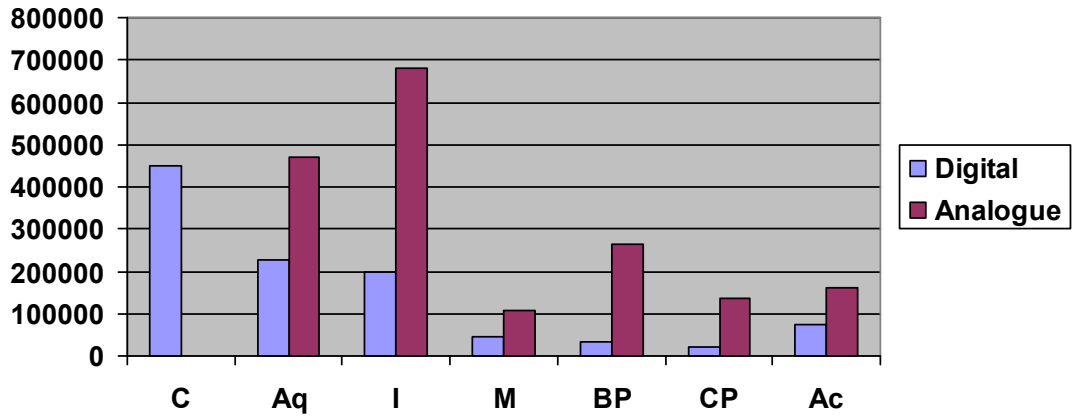
the Case Study. These are the total project costs allocated into the LIFE model based upon the functions identified in the workflow section.

Table 39 - Summary of total project costs

	C	Aq	I	M	BP	CP	Ac	Total
Digital	£448,456	£228,781	£196,820	£44,171	£34,813	£19,625	£72,921	**£1,045,587**
Analogue	0	£471,199	£679,466	£107,474	£265,273	£137,565	£159,726	**£1,820,702**

The following graphical representation of the total project cost is added to illustrate the costs visually to give an "at a glance" view of where the costs differ between an analogue object and a digital object.

Figure 13 - Total project costs



This table divides the total project costs by the number of entities to give a total per entity cost

Table 40 - Project Costs by Entity

	Total Project cost	Total number of entities	Per entity cost
Digital	£1,045,587	190,256	£5.50
Analogue	£1,820,702	399,000	£4.60

5.9.2 Per Entity Cost Split by LIFE Stage

This Table represents the total per-entity cost split by the LIFE stage level to show the cost per object split.

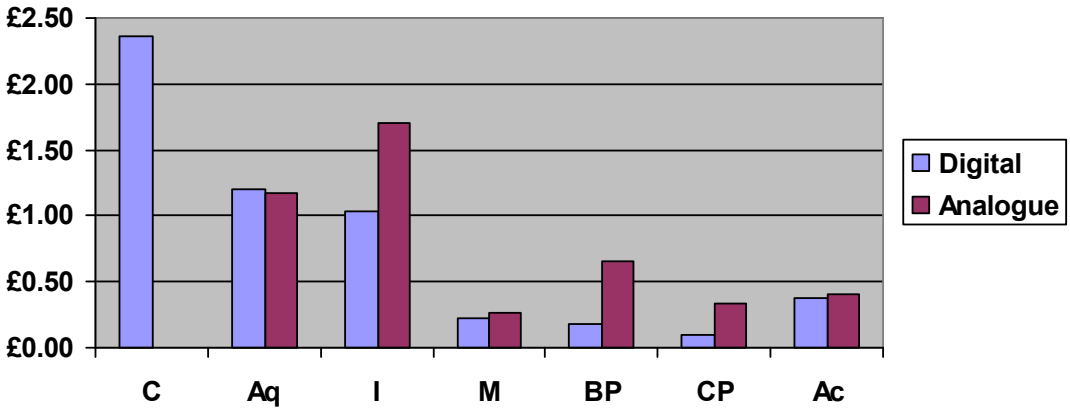
Table 41 - Per entity cost split by LIFE stage

	C	Aq	I	M	BP	CP	Ac	Total

Digital	£2.40	£1.20	£1.00	£0.20	£0.20	£0.10	£0.40	**£5.50**
Analogue	£0.00	£1.20	£1.70	£0.30	£0.70	£0.30	£0.40	**£4.60**

Figure **14** is a graphical representation of the total entity cost by LIFE stage.

Figure 14 - Per entity cost split by LIFE stage



5.9.3 Trend Analysis of Creation Cost plus Page and Article Level Costs

When figures at the object and entity level started to be finalised, it became clear that a significant proportion of the digital object lifecycle cost was occurring in the Creation Stage. In fact if the cost of creation for digital materials is excluded from the total cost, then the per-entity cost drops to £3.12 compared to £4.56 for analogue materials.

Given this scenario, the LIFE team decided that a comparison of the cost of creation might be a useful addition to the cost analysis just to see whether the cost to create digitised content has come down since the completion of the Burney digitisation project.

Secondary to this requirement, it was also thought that page-level and article-level information would be a useful statistic in terms of cost analyses. Unfortunately in the analogue procedures for the ingest of newspapers, page- and article-level information is not kept so JISC1 Newspaper digitisation figures have been used here for comparison. It is therefore important to note that this section is comparing two digital collections to show both a reduction in creation costs plus a comparison of per-page and per-article information.

This Table is a summary of costs between the JISC1 newspaper digitisation creation-only (scanning and OCR) costs compared to the Burney digitisation creation-only (scanning and OCR) costs. This figure is more relevant to this Case Study as the largest proportion of cost within the digital object cost resides in the creation of digital files and the required Optical character recognition.

Table 42 - Digital Creation cost comparison between JISC1 and Burney

Description	Creation cost (scan and OCR)	Number of pages	Cost per page	Reduction in cost from Burney to JISC1 per page
Burney 18th C	£448,456	£916,652	£0.50	-35%
JISC1 19th C			£0.30	

What this clearly shows is a reduction in processing costs for scanning and optical character recognition over the course of the last three years. Burney creation costs for scanning were high due to the experimental nature of the project. The JISC1 figures show that both technology improvement and the experience of dealing with large digitisation projects have brought the costs down by some 35%. This shows that digitisation costs are continuing to fall for large-scale projects and starts to make digitisation more cost-effective when compared to an analogue workflow.

Table 43 – Per-page Comparison between JISC1 and Burney

Description	Total project cost	Number of pages	Cost per page	Reduction in cost
Burney 18th C	£1,045,587	£916,652	£1.10	-11%
JISC1 19th C			£1.00	

At page level, Table 43 shows that the throughput rate for digitised pages (both projects ran for similar lengths of time) has grown by 124%, whilst the total overall project cost has dropped per page by 11%.

Table 44 – Per-article Comparison between JISC1 and Burney

Description	Total project cost	Number articles	Cost per article	Reduction in cost
Burney 18th C	£1,045,587	£1,878,234	£0.60	-38%
JISC1 19th C			£0.40	

At article level (Table 44) the costs for JISC1 have dropped by 38%. However the average number of articles from 18th Century Newspapers to 19th Century Newspapers means that the cost per article level figure was always likely to represent a large reduction.

5.10 Conclusions

The level of analysis that was required to identify the costs associated with an analogue collection was not insignificant. Considerable effort has been required to produce detailed business analysis of the functions and costs that a large analogue collection entails. This analysis provided a strong challenge to the methodology of the LIFE approach which had until this time been used

solely for digital collections. This has led to a much tighter definition of the methodology and the steps that are mandatory to produce consistent results.

This approach and methodology is best summarised as:

1. Interviews with project staff
2. Production of workflow diagrams
3. Identification of lifecycle functions
4. Spreadsheet work - financial analysis of lifecycle functions (staff costs etc.)
5. Allocation of lifecycle functions into the LIFE model
6. Creation of Table mapping the lifecycle functions to LIFE stage and element levels
7. Total cost analysis and LIFE stage analysis
8. Conclusions.

The LIFE Model terminology was a better-than-expected fit for analysis across digital and analogue collections. Only two changes were required to the stage level definitions, and none whatsoever was required at element level. At sub-element level it was felt that the fit was probably less accurate and that terms specific to the collection being analysed would be used. The best way to think of these would be to look at the descriptions of lifecycle functions in section 5.1.

The two changes at stage level that were required were the use of the phrase 'Book Storage Provision' instead of 'Bit-Stream Preservation' and 'Conservation Procedures' instead of 'Content Preservation'. Both changes were well accepted by the analogue project teams and (as described in the Case Study) were felt to clearly represent the difference between analogue and digital objects.

The addition of the workflow diagrams is an essential addition to all future analysis. The ability to plot the workflow to the LIFE Model will help in any future studies.

The model has been shown to work effectively across these analogue and digital collections. This is expected to become a very useful way for libraries to compare the costs associated with both types of object. This may lead to future collection management decision making in the areas of Acquisition, Storage or Preservation.

Although the model has been effective in the identification of functions and costs associated with analogue or digital collections, it has proven more challenging to compare one format against the other. It would be unfair to say that the headline costs prove that analogue lifecycle costs are cheaper than digital due to the differences between the two collections. The main difference is that the digital collection is a project that is static in size whereas the analogue collection continues to grow. The next logical stage in this type of analysis, therefore, is to compare two collections that are either static or growing. Given the work that has been done in LIFE^2, this should be a much easier objective for any future research.

The addition of Creation as a LIFE stage has introduced an issue for comparative analysis. National libraries receive much of their collection through Legal Deposit and so incur no charge for Creation or Purchase. However, as this analysis shows, Creation costs for Burney digitisation account for 42% of the total project cost. This fact means that for true comparison for National Libraries, these costs may need to be removed.

The realisation that Creation costs were such a high part of the lifecycle cost has led to a small piece of additional research by the LIFE^2 team. By comparing the creation costs from Burney digitisation with a more recent project, LIFE estimates that large-scale digitisation and OCR costs are dropping at around 12% over a three years period.

In the analogue collections, most costs appear in the Ingest and Acquisition stages. This is due mainly to the manual procedures that are carried out in the day-to-day operation of a paper-based repository. The lifecycle costs for analogue objects are dominated by labour costs and it is in the areas of Acquisition and Ingest that they are most prevalent.

The development of the LIFE spreadsheets from LIFE[1] to LIFE[2] has helped to capture costs for digital and analogue collections more effectively and consistently. Much work has gone into refining the spreadsheet calculations and these are expected to become a key part in the development of a software tool to cost future digital and analogue preservation.

5.10.1 Closing Comment

The headline conclusion is that the LIFE v1.1 Model has been an effective tool in enabling the evaluation of both analogue and digital lifecycle costs. Additionally a number of issues and outstanding questions have required careful consideration and it is hoped that this research has overcome most of the issues of comparative analysis for analogue and digital objects.

The aim of this Case Study was to see whether the lifecycle cost of analogue objects could be identified and mapped against that of digital collections and this has been done. It was not a Case Study to determine which method is cheaper or more expensive, although a by-product of the research is that it is possible to see the results of the costs side by side.

The only clarification that the LIFE team think is important to make is that the creation cost for digital material has had a major impact on the total lifecycle cost of a digital entity. For analogue materials, because of the legal deposit situation, no creation costs are counted, but the team knows that there are of course creation costs incurred in other areas outside the institutional responsibility outlined here.

So for final costing purposes, the team feels that the most realistic comparison would be the digital object cost *minus* creation cost versus the equivalent analogue object cost which results in the per-entity cost below.

Table 45 - Total Per-entity Cost Minus Creation Cost

	C	Aq	I	M	BP	CP	Ac	Total
Digital		£1.20	£1.00	£0.20	£0.20	£0.10	£0.40	**£3.10**
Analogue		£1.20	£1.70	£0.30	£0.70	£0.30	£0.40	**£4.60**

6 FINDINGS & CONCLUSIONS

6.1 Purpose of this Section

This section outlines the overall findings and conclusions from the project as discussed throughout this Report. It is broken up as follows:

- Economic Evaluation of LIFE
- LIFE Model v2
- Generic Preservation Model v1.1
- Institutional Repositories Case Studies
- British Library Newspaper Case Study
- Conclusions for Digital and Analogue Comparison
- Overall Conclusions

6.1.1 Supporting Documents

The Case Study findings and conclusions are drawn from two key areas. First, the Case Study write-ups (Section 4 for Institutional Repositories and Section 5 for Newspapers) contain detail on the process behind mapping to the LIFE Model. Second, all of the costs themselves are available in the spreadsheets which can be downloaded from the LIFE website (www.life.ac.uk).

Each spreadsheet contains a summary sheet, as well as a full lifecycle breakdown.

The workflows for Newspapers and SHERPA DP Case Studies are also available for download as individual files.

6.1.2 A Note on Costs

As noted in the Methodology, when examining the costing examples throughout the report number of points should be observed.

While the LIFE team do calculate exact costings with pounds and pence, a more meaningful way of looking at the lifecycle costs is through the graphs. As is discussed throughout the Case Studies, the final costs are accurate, but are very collection-dependent, as well as being based on certain assumptions that are not homogonous across other institutions. The graphs (rather than the costing tables) allow for a more general picture of the lifecycle to be observed.

It can be misleading to take the costing in the spreadsheets as absolute. As noted throughout the report, for certain Case Studies the costings should be regarded as illustrative rather than absolute. For reference, the spreadsheets do give exact costing calculations with no alterations to the figures. However, the per-entity cost tables in this report use figures that are rounded up by at least one significant figure.

6. Findings & Conclusions
Economic Evaluation
LIFE Model v2
Generic Preservation Model
SHERPA DP
SHERPA-LEAP
Newspapers
Case Study Conclusions
Review
Concluding Comment

6.2 Economic Evaluation of LIFE

When the first phase of LIFE was completed, one of the key elements that the team wanted to work on for LIFE^2 was a review of the economic approach used. Professor Bo-Christer Björk from Hanken, the Swedish School of Economics and Business Administration, was brought on board to complete a full independent review to the LIFE approach.

The report largely validated the approach taken by the LIFE team. At the same time, it provided a number of recommendations to steer the second phase of the project in the right direction on key economic issues such as the use of discounting, the role of inflation and costs outside of the lifecycle. The review recommended that all calculations were done using real-term, inflation-adjusted costs. It also recommended that no discounting should be applied. The recommendations are summarised in Section 4.3 (page 39), and the full independent review is available from the LIFE Website[42].

6. Findings & Conclusions
Economic Evaluation
LIFE Model v2
Generic Preservation Model
SHERPA DP
SHERPA-LEAP
Newspapers
Case Study Conclusions
Review
Concluding Comment

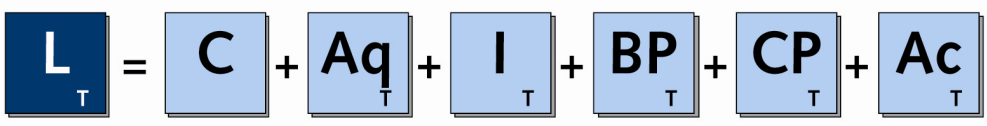
6.3 The LIFE Model v2

The LIFE Model provides a view onto the typical processes applied to digital objects throughout their lifecycle by an organisation acting as the custodian of those objects. The processes are loosely organised in a chronological order, from their creation through to eventual access. It should be noted however that processes can, in practice, overlap with each other or be executed in a different order. The Model aims to capture common processes found in most digital lifecycles. While some processes may not be applicable to all lifecycles, the intention is to provide meaningful placeholders for the majority of typical lifecycle processes.

The LIFE Model has gone through several reviews since it's inception in LIFE^1. The first version of the model used throughout the first phase of the LIFE Project ('LIFE Model v1') was used as the starting point for a thorough review that ended with the production of a working update of the Model that was used on all of the Case Studies in this second phase of the project ('LIFE Model v1.1').

A final review of the Model based on feedback from the LIFE2 Conference, the digital preservation community, early adopters of the LIFE work, and feedback from the Case Studies led the term to produce the final updated version of the model for this phase of the project ('LIFE Model v2'). It is this version of the model that is explained in Section 3.4 of this report (page 17) and is summaries in Figure 15 and Figure 16 below.

Figure 15 - Stages of the LIFE Model v2



[42] Björk, B.-C. (2007) *Economic evaluation of LIFE methodology.* Research report. LIFE Project, London, UK. Available from: http://eprints.ucl.ac.uk/7684/

Figure 16 - The LIFE Model v2

Lifecycle Stage	Creation or Purchase [43]		Acquisition	Ingest	Bit-stream Preservation	Content Preservation	Access
Lifecycle Elements	....		Selection	Quality Assurance	Repository Administration	Preservation Watch	Access Provision
	....		Submission Agreement	Metadata	Storage Provision	Preservation Planning	Access Control
	....		IPR & Licensing	Deposit	Refreshment	Preservation Action	User Support
	....		Ordering & Invoicing	Holdings Update	Backup	Re-ingest	
	....		Obtaining	Reference Linking	Inspection	Disposal	
	....		Check-in				

6.4 Generic Preservation Model (GPM) v1.1

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[44]. There are a number of isolated examples of preservation action but there is still very little information available. While it is over two years since the completion of the first phase of LIFE , there are still few details 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 Generic Preservation Model aims to both identify and estimate the cost of 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.

Section 3.5 (page 34) summarises the update to the preservation model with an accompanying spreadsheet. This model allows institutions to estimate potential digital preservation costs for their collections. One of the outcomes from the Case Studies, is that clearly, it is still very difficult to gain actual (and accurate) digital preservation activity costs. Unfortunately, this

[43] This stage may be beyond the scope of some costing activities. Creation may occur outside the view of the costing institution. It should therefore be considered to be optional. Where considered within scope, elements will need to be tailored to the specific lifecycle case in question.
[44] See Cedars Project, Research Review, LIFE[1], http://eprints.ucl.ac.uk/1856/1/review.pdf

estimative GPM still seems to be the only way to provide an indication of what some of the long-term costs of preservation might be.

This predictive modelling of the preservation costs is an area that needs to be further worked on within the digital preservation community. As outlined in the final section of this report on areas of future work, by developing predictive models to populate the other Stages of the lifecycle the LIFE team could gain a much larger sample of costs than the Case Study approach allows for. This might well be the best way forward to further develop this difficult and complex area.

6.5 SHERPA DP Case Study

The results for SHERPA DP are broken down into the lifecycles costs for Year 1, the total costs for the first 5 years of the lifecycle and finally the overall costs over a 10-year period.

6.5.1 Lifecycle Costs in Year 1

Figure 17 - SHERPA DP Lifecycle Costs (Year 1)

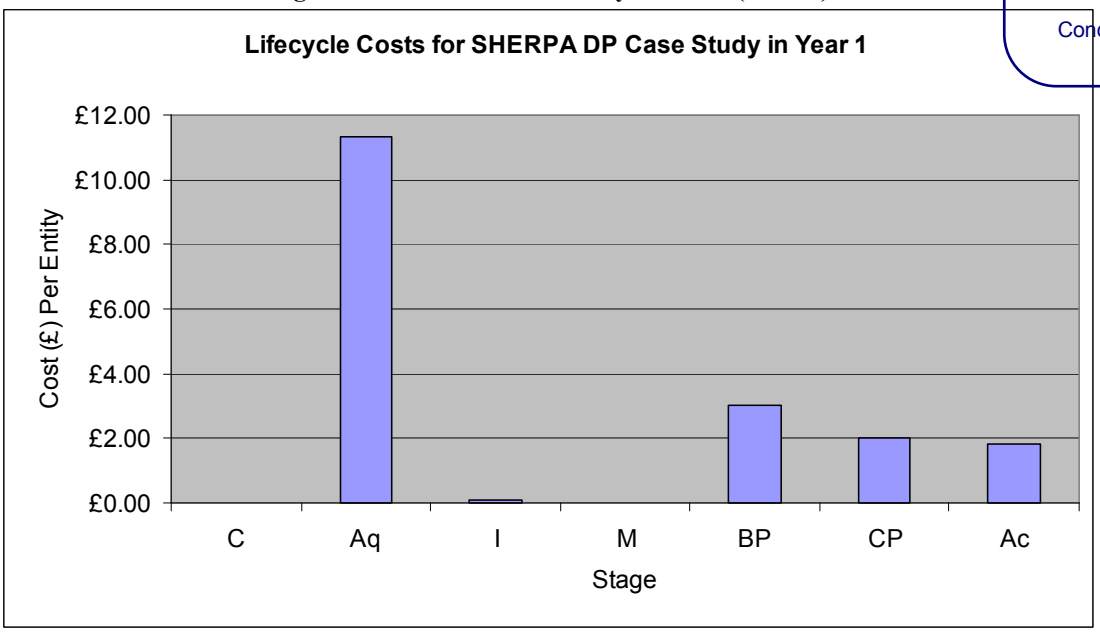


Table 46 - SHERPA DP Lifecycle Costs Per Entity (Year 1)

Stage	C	Aq	I	M	BP	CP	Ac	Total
Cost	£0.00	£11.40	£0.10	£0.00	£3.00	£2.00	£1.80	£18.40

Table 47 - SHERPA DP Total Lifecycle Costs (Year 1)

Stage	C	Aq	I	M	BP	CP	Ac	Total
Cost	£0	£74,050	£763	£0	£19,848	£13,233	£11,901	£119,801

6.5.2 Total Lifecycle Costs over 5 Years

Figure 18 - SHERPA DP Lifecycle Costs (Total over 5 Years)

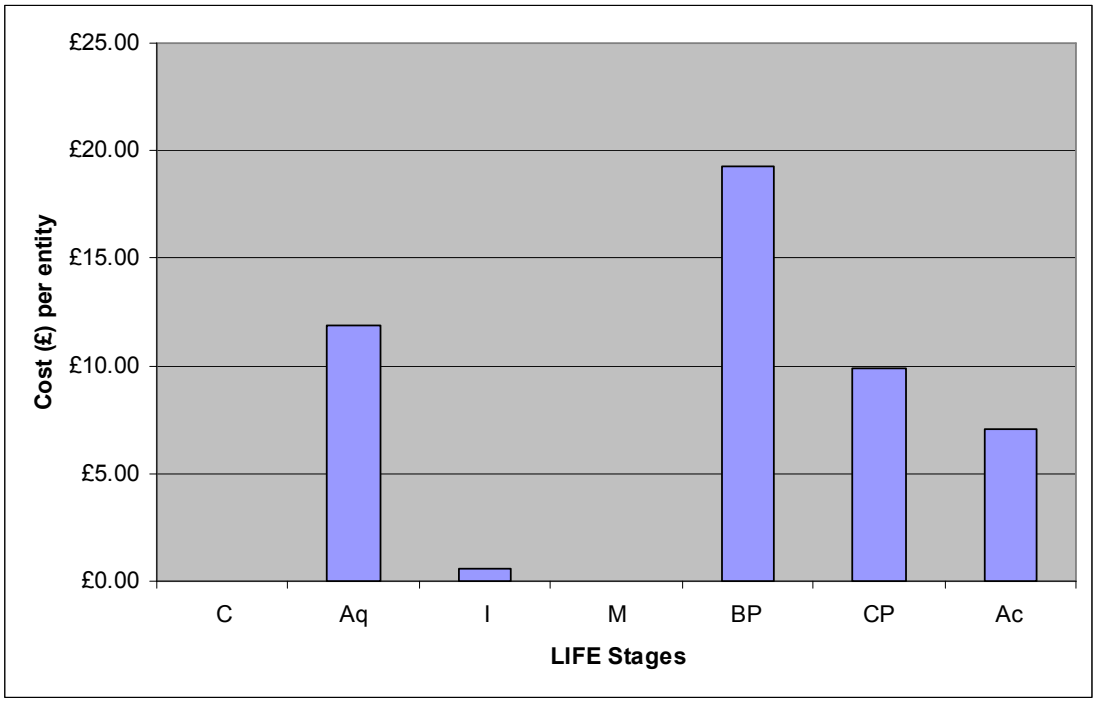


Table 48 - SHERPA DP Lifecycle Costs Per Entity (Total for 5 Years)

Stage	C	Aq	I	M	BP	CP	Ac	Total
Cost	£0.00	£11.90	£0.60	£0.00	£19.30	£9.90	£7.00	£48.70

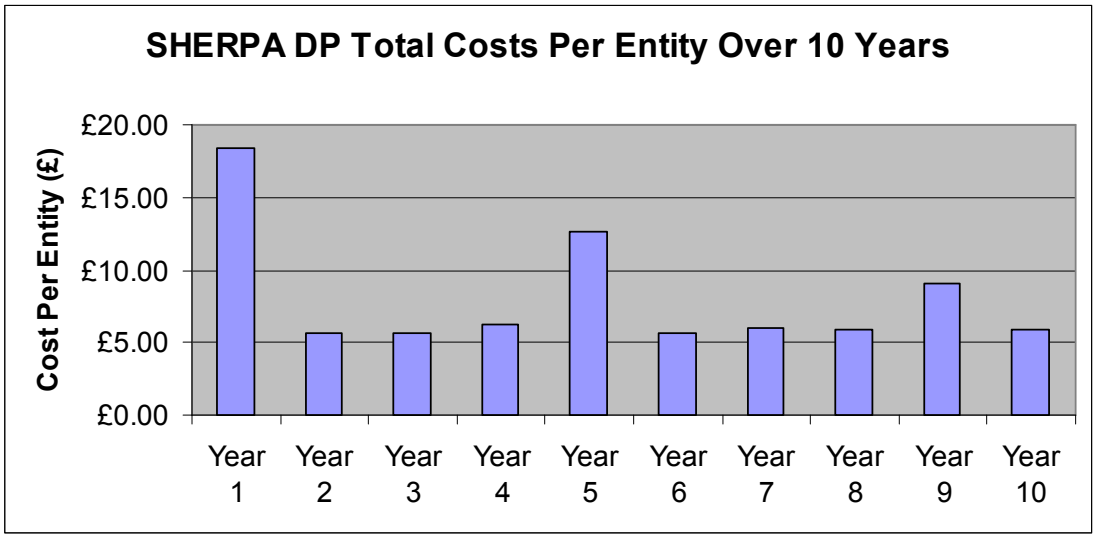
Table 49 - SHERPA DP Lifecycle Costs (Total for 5 Years)

Stage	C	Aq	I	M	BP	CP	Ac	Total
Cost	£0	£77,510	£3,841	£0	£125,870	£64,615	£45,875	£317,711

6.5.3 Costs over 10 Years

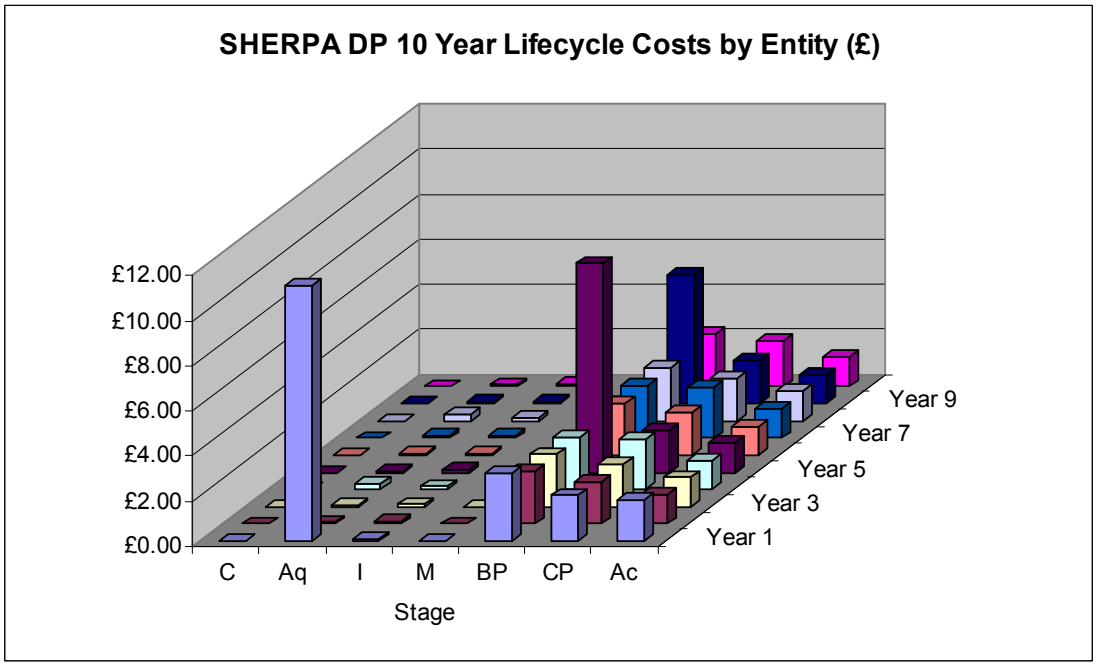
Figure 19 and Figure 20 highlight some of the lifecycle cost trends for SHERPA DP over a 10-year period. The total lifecycle costs (Figure 19) show that there are spikes in activity in Years 1, 5 and 9.

Figure 19 - Total Costs over 10 Year period



When looking at these same costs broken down by Stage (Figure 20), it is clear that the spikes in costs are due to the high Acquisition costs for Year 1 and the spikes in Bit-stream Preservation activity for years 5 and 9.

Figure 20 - 10 Year Lifecycle Costs per Entity



6.5.4 Strategic Findings

The key finding for this Case Study was that the costs did not vary greatly for differing quantities, as a largely-automated process has been established. There were 6,526 objects harvested as part of the process for SHERPA DP, giving the overall costs highlighted in Table 30.

Table 50 - Summary of Total Costs from SHERPA DP Case Study

	Total Cost	Cost per Object	Annual Cost per Object
Year 1	£119,801	£18.40	£18.30
Years 1-5	£317,711	£48.70	£9.70
Years 1-10	£530,515	£81.30	£8.10

As established in the Case Study write-up (Section 4), there were no costs for Creation or Purchase. Acquisition costs were mostly for the development of the OAI-PMH tool and for integrating the harvester with the AHDS repository. Ingest costs were low, since quality assurance was the responsibility of the source repositories: scheduled harvesting using OAI-PMH led to file format characterisation being automated using DROID.

The largest cost area was in Bit-stream Preservation, since this included staff elements for system administration and technology monitoring, as well as for storage provision.

As with the other Case Studies, Preservation Action was a particularly hard part of Content Preservation to cost, while Preservation Planning and Technology Watch are more consistent across time.

6.6 SHERPA-LEAP Findings

The findings for this Case Study are used to compare the lifecycle costs for the three repositories over one year. As demonstrated in the Case Study discussion in the previous section, the costs indicated here should be regarded as illustrative, rather than absolute.

As shown in Table 51, the Year 1 costs per object at Goldsmiths are £31.48, at Royal Holloway £23.13, and at UCL £15.98.

Table 51 - Overall Costs for SHERPA-LEAP Repositories

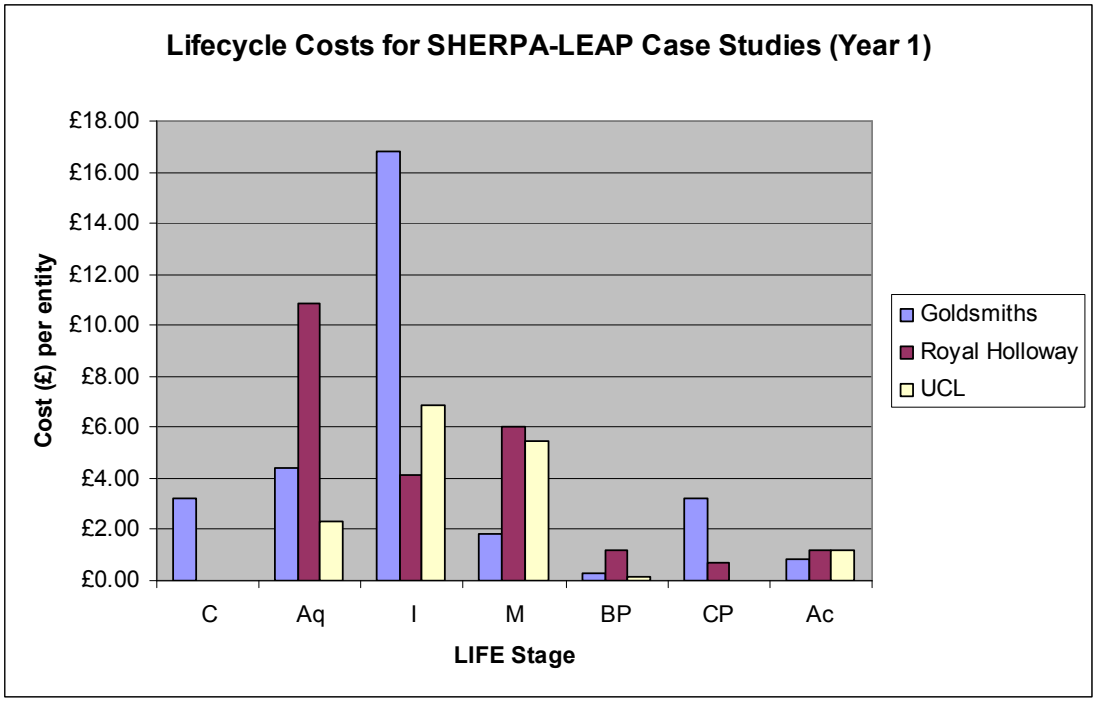
	Year 1	Year 5	Year 10
Goldsmiths	£31.50	£32.00	£32.20
Royal Holloway	£23.10	£23.60	£23.90
UCL	£16.00	£16.50	£16.70

6.6.1 SHERPA-LEAP Per Entity Repository Lifecycle Costs

Table 52 - Repository Lifecycle Costs Per Entity (Year 1)

Stages	C	Aq	I	M	BP	CP	Ac	Total
Goldsmiths	£3.20	£4.40	£16.80	£1.80	£0.30	£3.20	£0.90	**£30.60**
Royal Holloway	£0.00	£10.90	£4.10	£6.00	£1.20	£0.70	£1.20	**£24.10**
UCL	£0.00	£2.30	£6.90	£5.50	£0.10	£0.00	£1.20	**£16.00**

Figure 21 - Repository Lifecycle Costs (Year 1)



6.7 Conclusions for Institutional Repositories

6.7.1 Lifecycle Comments

All of the Institutional Case Studies brought up the issue of Metadata being a separate cost area in its own stage, as it was in the LIFE Model v1.1 (Figure 6, on page 31) which was used for the Case Studies. As a result of this (and other feedback), the Metadata processes have been integrated with the rest of the Model (as can be seen in v2 of the Model - Figure 4 on page 18). Having Metadata spread across the lifecycle better represents its part in the lifecycle of digital objects.

In all of the Case Studies, it was felt that identifying the activities on the Stage, Element and Sub-Element level was a valuable exercise to go through. Each of the Case Studies identified different cost spikes depending on the nature of their collections. The costs for each of the three Eprints repositories for example varied according the nature of their collections as well as the maturity of their repositories.

As a largely automated service, SHERPA DP could offer significant cost savings with increased quantity. It will be interesting to see how this develops with the next phase of SHERPA DP, which will test larger ingest actions. This will allow CeRch to calculate a new unit cost over time. It is hoped that this will validate the predictive lowering of unit costs with increased volume. These new costs will help in SHERPA DP's (and CeRch's) efforts to demonstrate the viability of a third-party preservation service.

One of the key organisational benefits was that the LIFE Model gave a meaningful structure to a costing exercise. Overall each of the Case Studies mapped well to the LIFE model, even though they were each very different from previous LIFE Case Studies run in 2006.

For each of the Case Studies, there was generally little issue with mapping the collections to the Stage and Element Level (the one or two exceptions listed in the Case Study write-ups). However, it was not always straightforward identifying the costs at the sub-element level. As with the previous LIFE[1] Case Studies, some institutions still encounter considerable difficulty in identifying specific costs at this level of detail.

As detailed in the SHERPA-LEAP Case Study discussion, the costs indicated in the results section should be regarded as illustrative, rather than absolute. This is due to several factors, among them:

- the patchwork nature of repository funding
- the need to base some costs on assumptions about future growth, expenditure, and preservation requirements
- differences in costing methods and interpretations of the LIFE v1.1 model across repositories.

As was noted in a number of the Case Study write-ups, using a Case Study approach is a valuable way of populating the LIFE Model with real costs. However, it does have drawbacks. For example, as with the three institutional repository Case Studies, given the time restrictions in the project, it is only possible to provide a snapshot of lifecycle costings and processes.

However, each of the Case Studies has contributed in a number of ways:

- Case Study feedback has contributed to a final review of the LIFE Model (v2)
- Institutional Repository Case Studies can be viewed as a useful guide to both the costs and the processes involved in maintaining an HE repository
- Each of the three repositories used were at different stages of set-up and contained a range of objects

6.7.2 Costing Conclusions

The variations in costings between the institutions in the LEAP Case Study may be attributed to three factors. First, the caveats already listed above should be noted. Second, the narratives show staff on different grades, in differing proportions, working in the repositories. This naturally affects the costings. As the repositories become more stable, staff gradings and roles are likely to become regularised, and comparison across the HE community will become more informative. Third, the studies show that the fact that Goldsmiths handles a range of

complex digital materials within its institutional repository structure increases the average handling cost per object.

As with SHERPA DP, after year 1, the main lifecycle costs are those associated with preservation. For SHERPA-LEAP, Bit-stream Preservation costs are based on estimates, both of repository growth and in the technology marketplace. Content preservation will clearly bring costs for the partners in the future, but for the time being those costs are not easily predictable.

This is something that perhaps the Generic Preservation Model can help to answer once it has been further developed and tested. These differences across both the SHRERPA-LEAP repositories and the other Case Studies leads to questions as to whether or not LIFE can yet be used for inter-institutional comparison when the collections themselves are so variable. This is one of the reasons why the context of the Case Studies is so important, and it is critical not simply to take the lifecycle costs at face value.

There is also the question of time and resources taken up to identify these costs in the first place. Each of these Case Studies needed considerable time spent on them, both internally within the institutions in question and externally by the LIFE Team. It would be fair to say that each of the Case Studies took a much longer timeframe to develop that originally anticipated. This should not be underestimated by other institutions thinking of performing similar costing studies.

For each of the Case Studies the effort was certainly worthwhile, allowing the institutions to gain a greater understanding of their own costs and processes. As noted by the CeRch team in the SHERPA DP Case Study, it certainly helps to have a business requirement for determining costs, but applying the LIFE model to different institutional settings is recommended to all with an interest in digital curation and preservation.

6.7.3 Overall Strategic Conclusions

- The SHERPA DP Case Study shows that a 3rd-party preservation solution is possible for digital repositories in the UK
- As an automated service, SHERPA DP could offer significant cost savings when increased quantities of digital objects are processed
- For SHERPA DP, the largest cost area was in Bit-stream Preservation, since this included staff elements for system administration and technology monitoring, as well as provision for storage (including equipment renewal) and offsite duplicate storage
- The variation in costings identified in the SHERPA-LEAP case studies reveals that the rollout of institutional repositories in the UK is still in its infancy
- The costing figures prepared by the SHERPA-LEAP partners are not yet robust enough for definitive conclusions to be drawn; it would be too simplistic to make comparisons between institutional costs at this stage
- Digital preservation is yet to become embedded as a concept in the Higher Education community. This presents a major challenge in advocacy for the global digital preservation community

- In the SHERPA-LEAP Case Studies, it is suggested that after year 1 the main lifecycle costs are those associated with preservation. However, Bit-stream Preservation costs are based on estimates, both of repository growth and in the technology marketplace. Content Preservation will clearly bring costs for the partners in the future, but for the time being those costs are not easily predictable.

- The Goldsmiths Case Study suggests that higher costs may currently be associated with managing complex digital materials at an institutional level.

6.8 British Library Newspapers

The key finding for this Case Study is that the LIFE Model has been an effective tool in enabling the evaluation and comparison of analogue and digital lifecycle costs. Certainly as a result of the Case Study, the team now has comparable costs for analogue and digital newspaper collections. However, it should be noted that the costs should not be taken out of context. When comparing analogue and digital lifecycles, each collection needs to be evaluated in its own right.

The findings for this Case Study are split into three sections:

- Total Project Costs
- Costs Per Entity (for Year 1 and as a 5-year total)
- Strategic Findings

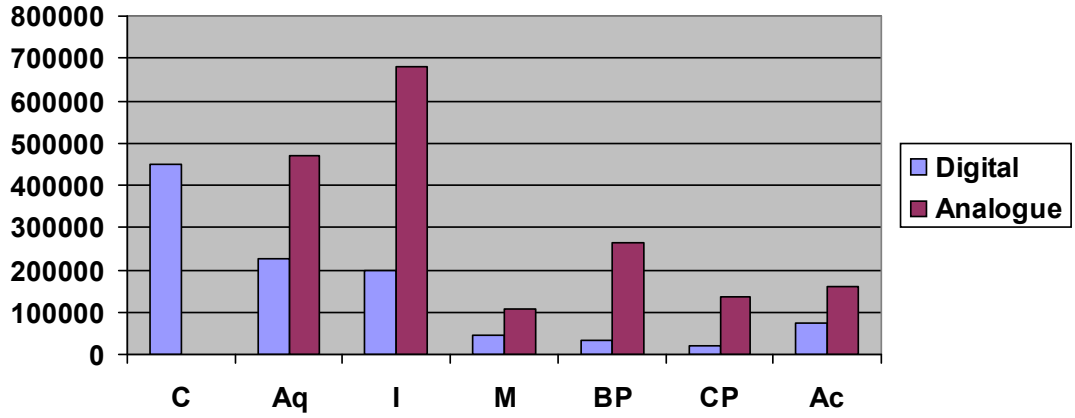
6.8.1 Total Project Costs

Table 53 summarises the total project costs across the newspaper lifecycle, while Figure 22 represents these costs graphically. This visual snapshot of the costs can be used to view where the costs differ between an analogue object and a digital object.

Table 53 - Summary of Total Project Costs

	C	Aq	I	M	BP	CP	Ac	Total
Digital	£448,456	£228,781	£196,820	£44,171	£34,813	£19,625	£72,921	**£1,045,587**
Analogue	£0	£471,199	£679,466	£107,474	£265,273	£137,565	£159,726	**£1,820,702**

Figure 22 - Total project costs (£)

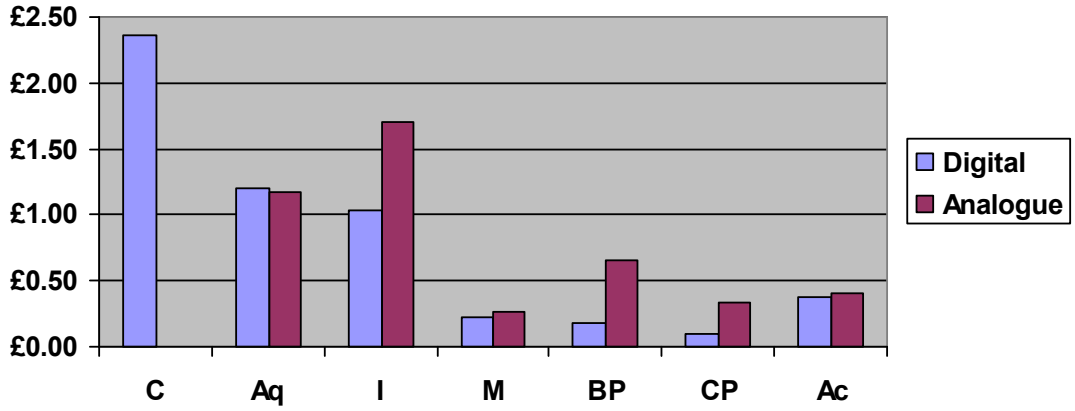


6.8.2 Per Entity Costs – Year 1

Table 54 - Total per entity cost minus Creation cost (Year 1)

	C	Aq	I	M	BP	CP	Ac	Total
Digital		£1.20	£0.90	£0.20	£0.30	£0.10	£0.40	**£3.10**
Analogue		£1.20	£1.50	£0.30	£0.90	£0.30	£0.40	**£4.60**

Figure 23 - Per entity cost split by LIFE stage



6.8.3 Per Entity Costs – 5 Year Total

Table 55 - Total per entity cost minus Creation cost (5 Year Total)

	C	Aq	I	M	BP	CP	Ac	Total
Digital	£2.40	£1.20	£0.90	£0.20	£0.30	£0.10	£0.40	**£5.50**
Analogue	£0.00	£1.20	£1.50	£0.30	£0.90	£0.30	£0.40	**£4.60**

6.8.4 Conclusions for Digital and Analogue Comparisons

The LIFE Model and associated methodology provided a useful way of comparing analogue and digital lifecycles. The resulting figures were considered to be a useful indication, if not an exact representation and comparison, of analogue and digital costs.

Other key conclusions:

- Comparison between analogue and digital lifecycles is complex and requires a great deal of effort, both to develop useful mappings and to generate accurate costs
- Analysing activity retrospectively was challenging. Costing activity as it occurs would be expected to be considerably more straightforward
- The application of the Model to an analogue lifecycle was workable, and the digital terminology used was understandable, and in most cases appropriate, for staff working with analogue collections
- A clear methodology and the use of workflow diagrams to illustrate complex processes considerably assisted the execution of the Case Study
- A number of the raw LIFE Stage costs calculated were surprisingly similar between the analogue and the digital lifecycles. Ingest and Bit Stream Preservation / Book Storage Provision were considerably higher for the analogue lifecycle.
- When creation costs are not taken into account (where a like with like comparison is not possible) the digital lifecycle was found to be marginally cheaper than an analogue lifecycle
- The analogue lifecycles which were examined are well-established and particularly efficient, but the digital lifecycles are relatively new and will see considerable streamlining and automation in the near future. Nonetheless, it appears that digital costs will before long be considerably lower than analogue costs. Trends in digitisation and wider lifecycle costs associated with newspaper content are discussed in Section 1

6.9 Overall Case Study Conclusions

Each of the Case Studies contributed in a variety of ways to the feedback process, however, one overall theme is clear - the development of the Case Studies is a complex process, requiring a great deal of time and effort.

Mapping an institution's processes to the LIFE Model can be a lengthy process, and gaining meaningful costs for these processes can rely heavily on activity-based costing, and in many cases estimation.

As a result of the process undergone for each of the Case Studies, a clear LIFE methodology for the Case Studies has been developed and refined. The workflow diagrams, which were used to illustrate complex processes, helped considerably in the execution and understanding of the Case Studies for Newspapers and SHERPA DP.

The differing nature of the Case Studies allows the team properly to test the LIFE Model with a variety of collections. However, it also means that it is critical that the context of each of the Case Studies is understood, and that the results are not simply taken at face value. Two key examples are worth highlighting here.

First, each of the three exemplars for the SHERPA-LEAP repository study had varying costs. It would be a mistake simply to label UCL Eprints more efficient or 'cheaper' because it had the lowest lifecycle costs. The value here is looking at *why* these costs are different and what lessons can be learned.

Second, it would be inaccurate to label the analogue lifecycle cheaper than a digital lifecycle as a result of the Newspapers Case Study. What this is study has shown is that this is a complex area, which clearly needs further investigation. What is important here is that the LIFE model is workable for both analogue and digital collections.

6.10 Section Review – Key Outputs from LIFE^2

This section has outlined all of the key outputs from this second phase of the LIFE Project. The LIFE approach has been validated by a full independent economic review and has successfully produced an updated lifecycle costing model (LIFE Model v2) and digital preservation costing model (GPM v1.1). The LIFE Model has been tested with three further Case Studies including institutional repositories (SHERPA-LEAP), digital preservation services (SHERPA DP) and a comparison of analogue and digital collections (British Library Newspapers) and these Case Studies have fed into both the LIFE Model and the LIFE Methodology.

The LIFE work has been successfully disseminated throughout the digital preservation and HE communities. Early adopters of the work include the Royal Danish Library, State Archives and the State and University Library, Denmark as well as the LIFE^2 Project partners. Furthermore, interest in the LIFE work has not been limited to the library and HE Sectors, with interest in LIFE expressed by local government, records offices, and private industry.

It should be noted that to gain a fuller understanding of any of these outputs, they should be viewed in the full context of the rest of the project. Each of the outputs are summarised here, with links to where the full discussion is available within this report.

- The **LIFE Methodology** gives an outline of the methodology used throughout the project and outlines how to use LIFE and get the most out of both the Model and the Case Study results.
  - 〉 See: Sections 2.7 (page 7) and 2.8 (page 12)

- **Aims of Digital Preservation Costing** highlights some of the different approaches that an organisation can take to costing activities and how the LIFE approach fits in with these options.
  〉 See: Section 3.3 (page 15)

- **LIFE Model v1.1** is a working version of the LIFE Model used for the LIFE^2 Case Studies in order to gain feedback on the direction which the model should take. This version was used as a basis for the final model update (v2).

- **LIFE Model v2** outlines a fully-revised lifecycle model taking into account feedback from user groups, the Case Studies and the wider digital preservation community.
  〉 See: Section 3.4 (page 17)

- **Generic Preservation Model** (GPM) summarises the update to the preservation model with an accompanying spreadsheet. This model allows institutions to estimate potential digital preservation costs for their collections. The GPM fits into the updated LIFE Model.
  〉 See: Section 3.5 (page 34)

- **An Economic Evaluation of LIFE** outlines the report written by economist Bo-Christer Björk on the approach used for both the first and second phases of LIFE. This independent review validates the LIFE approach for lifecycle costing.
  〉 See: Section 3.2 (page 14)

- **SHERPA DP Case Study** outlines the mapping of the repository services that CeRch provides to the LIFE Model.
  〉 See: Section 4.3 (page 39)

- The **SHERPA-LEAP Case Study** maps three very different HE repositories to the LIFE Model. Goldsmiths University of London, Royal Holloway University of London and UCL (University College London) each provide exemplars of varying collections. Each institution's repository is at a different stage of development.
  〉 See: Section 4.4 (page 54)

- The **Newspapers Case Study** successfully maps both analogue and digital newspaper collections to the LIFE Model. This success means that LIFE could be developed into a fully-compatible predictive tool across both analogue and digital collections, allowing for comparison both throughout the lifecycles of a collection and across different types of collections.
  〉 See: Section 1 (page 75)

6.11 Concluding Statement – from LIFE[1] to LIFE[2] to LIFE[3]

The LIFE Project has made a major contribution in highlighting the short- and long-term costs in this complex area of digital preservation. It has facilitated better planning, comparison and evaluation of digital lifecycles, allowing for a greater understanding for the safe-keeping of digital collections.

The first phase in 2005-06 produced a first lifecycle costing model for digital collections. It took the work used to map traditional (analogue) collections and moved it into the digital realm, with a specific emphasis on digital preservation.

The second phase (2007-08) has refined and updated this approach. Certainly this second phase has fulfilled its project aims, and can be viewed as another successful phase of the LIFE work. The LIFE approach has been validated by a full independent economic review, and an updated Model has been reviewed, refined, tested and fully updated. The Model has been used to successfully compare an analogue and digital collection. Yet, when looking at the bigger picture, this phase has highlighted more problems than solutions, asking more questions than perhaps it has answers to.

The digital preservation community still does not have an accurate picture of the costs of digital preservation. The Case Studies highlighted how difficult it is to gain real costs for the complete lifecycle. The Case Study method has been incredibly valuable for scenario building and testing the LIFE Model, but has this approach run its course?

The Model needs a much greater volume of raw cost data. This need cannot be supported via the Case Study approach, and the LIFE team would suggest that a software tool would provide the volume of costing data necessary to provide a truly accurate predictive model. This would allow LIFE to analyse a range of key research areas, such as the lifecycle of primary data. Such an approach would allow us to rapidly expand our knowledge-base for each stage of the lifecycle.

There is a clear need for a tool that can produce an accurate picture of a collection's entire lifecycle. Informed decisions need to be made for our ever-expanding digital collections, including choices between analogue and digital. Both National and HE Libraries need to be able to make these informed decisions in order to take a long-term view for the stewardship of their collections.

The development of a LIFE software costing tool would start to provide some answers to an area still very much filled with question marks.

7 FUTURE WORK AND LIFE[3]

7.1 Purpose of this Section

In both the first phase and this current phase of the LIFE Project, the team identified a number of key aspects of the work that would be useful to further develop in a third phase of the LIFE Project, LIFE[3].

7.2 The Next Phase – LIFE[3]

Feedback from the LIFE[2] Conference in June 2008, (in the form of both, discussion on the day, and questionnaire feedback by participants) indicated that there was a clear need for further development of the LIFE work. In particular, the development of a toolset to implement the LIFE Model was seen as a key direction for the work.

A third phase of the LIFE Project would focus on:

- developing a toolset to enable libraries/researchers/teachers/research funders/institutions to implement the LIFE costing models for lifecycle curation and long-term digital preservation;
- developing predictive models for each stage of the digital lifecycle;
- further integration of digital with analogue lifecycle costing;
- liaising with other institutions internationally in order to gain additional external Case Study feedback;
- building on the work of the LC-JISC Blue Ribbon Task Force on the Economic Sustainability of Digital Preservation, on which LIFE is represented by Dr Paul Ayris

A further phase of LIFE would deliver a practical costing toolset that will enable more effective planning for digital preservation by researchers, libraries, institutions, users and funders. It would facilitate the costing of existing digital processes as well as estimating the costs of handling new digital content.

A LIFE Planning Tool (LPT) would take as an input a simple profile of a digital collection or content stream. This tool would then automatically process this profile, and would estimate the costs for each lifecycle stage for the required timescale. The resulting output would provide invaluable information with which stakeholders could plan and resource for the acquisition, ingest, storage, preservation of and access to new content.

Questionnaire feedback from the conference leads the LIFE team to believe that there is considerable demand for such a tool within the HE and research communities.

7.3 Case Studies and Activity-based Costing

The LIFE Project has now run six full exemplar Case Studies with a variety of institutions. In each case it should not be underestimated the amount of time and resource necessary to do a full and thorough costing analysis. The project team would certainly encourage other institutions to consider adopting the LIFE approach; however, such an activity should be thoroughly planned out before embarking on such an exercise. There is also a sensitive path to tread between analysis and audit. Self-analysis for Case Studies rather than project staff analysing other institutions is preferable.

8 ACRONYMS

List of acronyms used throughout the LIFE Documentation. .

Ac	-	Access
AHDS	-	The Arts and Humanities Data Service
AIP	-	Archival Information Package
Aq	-	Acquisition
BCT	-	Base Cost of Testing (from the GPM)
BL	-	The British Library
BLE	-	Base Life Expectancy (from the GPM)
BL eIS	-	British Library e Information Systems
BMP	-	Bit-Mapped Graphics Format
BP	-	Bit-stream Preservation
C	-	Creation
CAMiLEON	-	Creative Archiving at Michigan and Leeds Emulating the Old on the New
CEDARS	-	CURL Exemplars in Digital Archives
CeRch	-	The Centre for e-Research
COA	-	Cost of Action (from the GPM)
CP	-	Content Preservation
CRS	-	Cost of a New Rendering Solution
DOM	-	Digital Object Management system at the BL
DLT	-	Digital Linear Tape
DTD	-	Document Type Definition
EISSN	-	Electronic International Standard Serial Number
ETA	-	Ending Proportion of Tool Availability
EU	-	European Union
FCLA	-	Florida Centre for Library Automation
FCM	-	Fixed Cost Migration (from the GPM)
FCX	-	File Format Complexity (from the GPM)
FE	-	Further Education (UK)
FP	-	EU Framework Programmes
FSF	-	Format Stabilisation Factor (from the GPM)
gb	-	Gigabyte(s)
GIF	-	Graphics Interchange Format
GPM	-	Generic Preservation Model
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	-	Lifecycle costing
LIFE	-	Lifecycle Information For E-Literature
LOCKSS	-	Lots of Copies Keeps Stuff Safe
M	-	Metadata
mb	-	Megabyte(s)
METS	-	Metadata Encoding and Transmission Standard
MLA	-	Museums, Libraries, and Archives
MIME	-	Multipurpose Internet Mail Extensions
N	-	Number
N/A	-	Not applicable
NESLI	-	National Electronic Site Licensing Initiative
OAIS	-	Open Archival Information System
OCR	-	Optical character recognition
OPAC	-	Online Public Access Catalogue
PCP	-	Per Object Cost of Preservation
PCX	-	a graphics file format for PCs
PDF	-	Portable Document Format
PDI	-	Preservation Description Information
PREMIS	-	PREservation Metadata Implementation Strategies
PLA	-	Planning: Action (from the GPM)
PLN	-	Planning: No Action (from the GPM)
PLoS	-	Public Library of Science
PNG	-	Portable Network Graphic
POC	-	Proportion of Collection (from the GPM)
POM	-	Proportion of Migration (from the GPM)
PON	-	Proportion of Normalisation (from the GPM)
PPA	-	Performing Preservation Action
PREMIS	-	Preservation Metadata Implementation Strategies
PTA	-	Proportion of Tool Availability
PUM	-	Per Unit Migration (from the GPM)
QA	-	Quality Assurance
QAA	-	Quality Assurance Actions
RAE	-	Research Assessment Exercise
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 (from the GPM)
TIFF	-	Tagged Image File Format
TLSS	-	Teaching and Learning Support Section, UCL Library Services
txt	-	ASCII text files
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

9 BIBLIOGRAPHY

Björk, B.-C. (2007) *Economic evaluation of LIFE methodology.* Research report. LIFE Project, London, UK. Available from: http://eprints.ucl.ac.uk/7684/

Beagrie, N. *et al.* (2008) *Keeping research data safe: a cost model and guidance for UK Universities.* JISC. Available from: http://www.ndk.cz/dokumenty/dlouhodoba-ochrana/keeping-research-data-safe-a-cost-model-and-guidance-for-uk-universities

Cedars Project. (2002) *Cedars Guide to Collection Management*. Available from http://www.leeds.ac.uk/cedars/guideto/collmanagement/guidetocolman.pdf

Finch, Lorraine. (2005) *Cellulose Acetate Microfilm Forum (CAMF).* Available from: http://www.bl.uk/services/npo/journal/3/camf.html

Knight, G. (2007a). *An investigation of file formats in use by SHERPA DP repositories.* Available from: http://www.sherpadp.org.uk/documents/wp61-fileformats.pdf

Knight, G. (2007b) *Recommendations to ensure the long-term preservation of digital objects stored by institutional repositories.* Available from: http://www.sherpadp.org.uk/documents/wp65-migration_review.pdf

Legal Deposit Act in the British Library. Available from: http://www.bl.uk/aboutus/stratpolprog/legaldep/

McLeod, R. , Wheatley, P. and Ayris, P. (2006) *Lifecycle information for e-literature: full report from the LIFE project.* Research report. LIFE Project, London, UK. Available from: http://eprints.ucl.ac.uk/1854/

SHERPA DP Project. (2006) *Technical Specification–Sherpa DP.* Available from: http://www.sherpadp.org.uk/documents/technical-specification.pdf

Nationaal Archief. (2005) *Costs of Digital Preservation.* Available from: http://www.digitaleduurzaamheid.nl/bibliotheek/docs/CoDPv1.pdf

Watson, J. (2005) *The LIFE project research review: mapping the landscape, riding a life cycle.* Literature review. London, UK. Available from: http://eprints.ucl.ac.uk/1856/1/review.pdf

Wheatley, P., Ayris, P., Davies, R., Mcleod, R. and Shenton, H. (2007) *The LIFE Model v1.1.* Discussion paper. LIFE Project, London, UK. Available from: http://eprints.ucl.ac.uk/4831/

10 OTHER PROJECTS

The following Projects were referred to in the LIFE^2 Report

10.1 Burney Digitisation Project

http://www.bl.uk/collections/early/burneydigitisation.html
The 17th-18th Century Burney Collection Newspapers Digitisation Project managed by Gale Cengage made all one million pages of the collection available to users in digital format.

10.2 CEDARS

http://www.leeds.ac.uk/cedars/
CEDARS is an acronym for CURL Exemplars in Digital Archives. CEDARS was a JISC funded project exploring themes of digital preservation. LIFE builds on some of the work carried out by CEDARS.

10.3 JISC Newspapers

http://www.bl.uk/collections/britishnewspapers1800to1900.html
The 19th Century British Library Newspaper Website, managed by Gale Cengage, was launched on 22 October 2007, with 1,000,000 pages of content.

10.4 LIFE^1

http://www.life.ac.uk/1/
The first phase of the LIFE project (LIFE^1) ran for 12 months ending with the highly successful LIFE Conference on 20th April 2006. LIFE is a project which aims to apply the theory of life cycle collection management to digital collections. The project provided practical information for all institutions who have an interest in collecting and preserving digital material.

10.5 SHERPA

http://www.sherpa.ac.uk/
SHERPA is investigating issues surrounding the future of scholarly communication. It is developing open-access institutional repositories in universities to facilitate the rapid and efficient worldwide dissemination of research.

10.6 SHERPA DP

http://www.sherpadp.org.uk
The SHERPA DP Project ran from 2005 to 2007, investigated a disaggregated service model and assign rights and responsibilities. The purpose of this project is to create a collaborative, shared preservation environment for the SHERPA project framed around the OAIS Reference Model.

10.7 SHERPA DP2

http://www.sherpadp.org.uk
SHERPA DP2 will extend the collaborative, shared preservation environment developed by the SHERPA DP project. This new project will build on that work by extending the implementation model to interact with repositories holding different and varied types of digital content and using a more diverse range of content management systems.

10.8 SHERPA-LEAP

http://www.sherpa-leap.ac.uk/
SHERPA-LEAP is a University of London (UoL) partnership, led by UCL, which has created open access institutional repositories at thirteen University of London institutions.

11 ACKNOWLEDGEMENTS

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

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

The LIFE Models
Thanks to the following for feedback and comments on the updated LIFE Model and updated Generic Preservation Model - Kevin Ashley (ULCC), Neil Beagrie, Bo-Christer Björk (Hanken, Swedish School of Economics and Business Administration), Ulla Bøgvad Kejser (Royal Library, Denmark), Frances Boyle (DPC), Peter Bright (British Library), Anders Bo Christensen (State Archives, Denmark), Birte Christensen-Dalsgaard (State and University Library, Denmark), Angela Dappert (BL), Adam Farquhar (British Library), Helen Hockx-Yu (BL), Birgit Nordsmark Henriksen (Royal Library, Denmark), Jan Nebber-Christensen (State Archives, Denmark), and Chris Rusbridge (DCC).

SHERPA-LEAP Case Study
Thanks to Rebecca Stockley (SHERPA-LEAP), Martin Moyle (UCL), Jacqueline Cooke (Goldsmiths), and Adrian Machiraju (Royal Holloway).

SHERPA DP Case Study
Thanks to Sheila Anderson, Steve Grace, and Gareth Knight at the Centre for e-Research (CeRch).

British Library Newspapers Case Study
Thanks Lucy Evans, Patrick Fleming, Richard Gibby, Ed King, Stephen Morgan, Deborah Novotny, Dawn Olney, and Bhavna Tailor from The British Library.

JISC
Thanks to Neil Grindley, Digital Preservation Programme Manager (JISC).

Project Support
Thanks to Chris Carrington at UCL for the website development and support, and to Erica McLaren for EPrints support at UCL. .