The LIFE Project

Bringing digital preservation to LIFE

Lifecycle Information for E-literature

A summary from the LIFE project Report Produced for the LIFE conference 20 April 2006

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

THE LIFE PROJECT TEAM

Rory McLeod

Project Manager and voluntary deposit of electronic publications case study author

Paul Wheatley

Web archiving case study author and generic preservation model development

Paul Ayris

e-journal case study author

Henry Girling

Project coordination

GOVERNANCE

The Project is being governed by an international Project board.
The full membership of the project board is as follows:

Dr Paul Ayris

Director of UCL Library Services and UCL Copyright Officer

Dr Richard Boulderstone

Director of e-Strategy and Information Systems, the British Library

Ms Eileen Fenton

Executive Director, Portico **http://www.portico.org/**

Dr Clive Field

Director, Scholarship and Collections, the British Library

Dr Erik Oltmans

Head of Acquisitions and Processing Division, Koninklijke Bibliotheek
(National library of the Netherlands)

Ms Helen Shenton

Head of Collection Care, the British Library

Executive summary

INTRODUCTION

The LIFE Project has developed a methodology to calculate the long-term costs and future requirements of the preservation of digital assets. LIFE has achieved this by analysing and comparing three different digital collections and by applying a lifecycle approach to each. From this work LIFE has identified a number of strategic issues and common needs.

The critical strategic issues are:

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

METHOD

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

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

COST

LIFE established that in the first year of a digital asset's existence;

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

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

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

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

PRESERVATION COSTS

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

OBSOLESCENCE WATCH

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

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

COLLABORATIVE UNDERSTANDING AND TOOL DEVELOPMENT

Differences between institutional workflow proved challenging in the LIFE Project, from acquisition and selection through to workflow and the allocation of costs. Most of these issues were overcome within the Lifecycle Model. However, a conclusion from LIFE has to be that in order to be successful at collaborative work there needs to be clarity about how your partner works. The greater the understanding of the differences and similarities, the higher the success ratio and the more realistic national standards and approaches become. LIFE strongly advocates this collaborative approach and would like to expand its experiences in this area to more accurately apply costs across a wider range of collections.

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

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

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

EXECUTIVE SUMMARY CONCLUSION

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

The following pages provide more detail on the research undertaken, the methodology and lifecycle cost breakdowns.

This document is a summary of the stages of comprehensive analysis undertaken for the full LIFE Project report on the accompanying CD-ROM.

Research review

In November 2005 a comprehensive review of existing lifecycle models and digital preservation techniques was undertaken. This was done in order to find a useable cost model that could be applied to the management of digital collections within a Library and HE/FE sector. This is a brief synopsis of the full research review.

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

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

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

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

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

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

The lifecycle methodology

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

The accuracy of the model is dependant on the quality and the quantity of the data inputs. By allocating a cost to as many relevant sections as possible, and by applying the Generic LIFE Preservation Model, a total lifecycle cost can be achieved. The model is;

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

L is the complete lifecycle cost over time 0 to **T**. Other categories are
Aq = Acquisition,
I = Ingest,
M = Metadata,
Ac = Access,
S = Storage,
P = Preservation

This model is intended to provide a broad-enough scope to be usefully applied to most digital collections, while providing enough specific elements to allow a detailed lifecycle breakdown.

The category and element break down used for LIFE are described below; full information on its implementation can be found in the final Report.

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)					

Case Studies in summary
(category comparisons)

The Case Studies form the backbone of this Project; they provide the data which has informed and refined the lifecycle model. Each of the Case Studies examined the everyday operations, processes and costs involved in their respective activities. The results of this research and recording activity were used to calculate and, where necessary, estimate the direct lifecycle costs. A sampling of these comparative lifecycle costs is shown below. A much greater level of detail can be found in the full Project Report.

The three chosen LIFE Case Studies are:

- VDEP – the voluntary deposited electronic publications collection of over 170,000 digital objects held at The British Library collected since January 2001.
- Web Archiving activities, The British Library's work as part of the United Kingdom Web Archiving Consortium (UKWAC) archiving 1000 sites per year.
- e-journals at UCL Library Services which holds subscriptions for 12,365 periodical titles providing services for over 19,000 students.

ACQUISITION

Case study name	Sub category	Year1	Year 10	10 year total cost by %
VDEP [1]	Electronic monographs	£0.00	£0.00	0%
	Electronic serials	£0.00	£0.00	0%
Web Archiving [2]		£107.67	£934.09	14%
UCL e-journals [3]		£374.09	£5159.07	98%

Acquisition provides contrasting views on the three Case Studies. Under VDEP there are no acquisition costs whereas, for UCL e-journals, acquisition represents the majority of the lifecycle costs. For Web Archiving selecting titles, seeking permission and web crawling contributes to costs in this category.

INGEST

Case study name	Sub category	Year1	Year 10	10 year total cost by %
VDEP	Electronic monographs	£1.70	£8.50	23%
	Electronic serials	£7.01	£220.13	13%
Web Archiving		£111.45	£1114.51	16%
UCL e-journals		£0.00	£0.00	0%

Significant ingest costs lie within differing areas of the VDEP and Web Archiving Case Studies. Intensive manual work on quality assurance represents around half of the Web Archiving cost. The VDEP ingest is primarily accounted for by deposit and the update of holdings information.

[1] VDEP costs are indicated by representative titles. Due to space limitations, costings devoted to handheld monographs and serials are not present in this Summary.

[2] Web Archiving costs are indicated by average cost per title, including costs for an average of just over 5 instances gathered per year.

[3] e-journals costs are indicated by average costs per title for the two 'Big Deal' Journal packages from a UK and continental European publisher for e-only delivery

METADATA

Case study name	Sub category	Year1	Year 10	10 year total cost by %
VDEP	Electronic monographs	£10.40	£13.80	37%
	Electronic serials	£20.05	£146.77	9%
Web Archiving		£4.25	£4.25	0.1%
UCL e-journals		£3.97	£3.97	0.1%

Metadata costs in the Web Archiving and e-journals Case Studies are minimal, where little metadata is recorded. This is an immature element of the Web Archiving process and is expected to increase in size as the activity develops. Manual, intensive cataloguing procedures within VDEP make the capture and recording of metadata a much more significant cost.

ACCESS

Case study name	Sub category	Year1	Year 10	10 year total cost by %
VDEP	Electronic monographs	–	–	–
	Electronic serials	–	–	–
Web Archiving		£4.03	£57.94	0.5%
UCL e-journals		£6.78	£79.53	1%

Both Web Archiving and e-journals feature relatively low access costs. VDEP as a system to test voluntary deposit offers no access.

STORAGE

Case study name	Sub category	Year1	Year 10	10 year total cost by %
VDEP	Electronic monographs	£4.01	£13.91	37%
	Electronic serials	£127.68	£1276.80	76%
Web Archiving		£53.91	£1078.17	8%
UCL e-journals		–	–	–

VDEP storage is based at the BL, contrasting with that provided by a third party for the Web Archiving Case Study. UCL does not store e-journals.

PRESERVATION

Case study name	Sub category	Year1	Year 10	10 year total cost by %
VDEP	Electronic monographs	£0.89	£1.45	4%
	Electronic serials	£10.68	£27.60	2%
Web Archiving		£425.50	£8509.91	62%
UCL e-journals		–	–	–

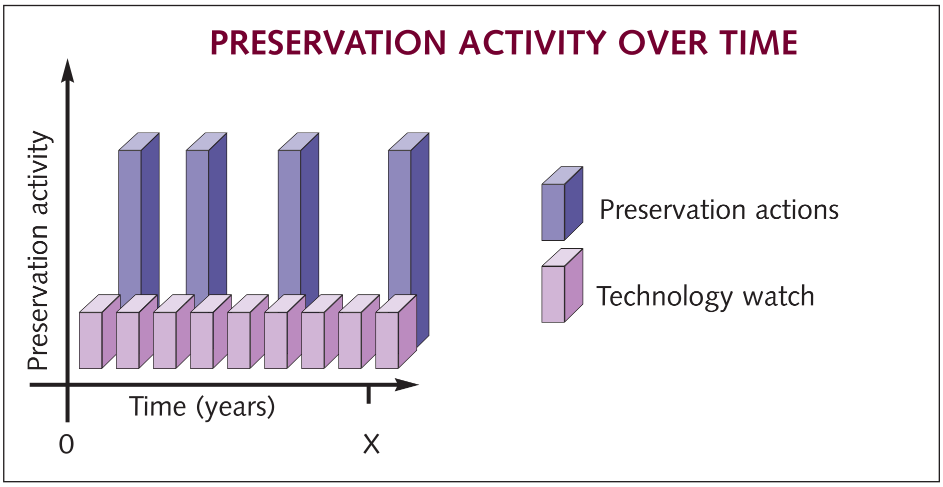
Preservation costs are estimated for both VDEP and Web Archiving, but provide an interesting contrast. The relatively low numbers of objects and small range of file types in VDEP results in quite modest preservation costs. Web Archiving represents the other extreme with an average web site instance containing over 6000 files, resulting in far higher preservation costs. Estimated preservation costs for UCL were not developed.

The Generic LIFE Preservation Model

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. There are a number of isolated examples of preservation action, but very little costing information has been recorded. Few details are available of either the breakdown of what the process might involve or of the costs of each of those elements for the large scale preservation of digital collections.

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

Because of the lack of historical figures, a strategy of estimation was employed. A mathematical model was developed to estimate costs for key areas of preservation activity. The model represents a series of costs over time, consisting of an annual technology watch and monitoring of content, along with spikes of preservation activity (as shown below).



Trends in areas such as tool development and the life expectancy of file formats were estimated and modelled. Inputs to the model for base costs like staffing were identified and defined. A process of review was performed to refine the model, and data from the Case Studies was employed to generate estimated preservation costs. A detailed explanation of the model and how it was constructed can be found in the full Report.

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

The following Table shows a sample of relative figures derived from the Preservation Model, utilising data from the Web Archiving Case Study.

The model suggests that the cost of performing preservation activities on an average instance of a captured web site (containing 6000 files) for a period of 20 years is around £80.

Conclusion

The LIFE Project has established that a lifecycle approach to cost is both applicable and useful for a range of digital collections. The three Case Studies show that variations in cost and workflow have been successfully captured within the LIFE Preservation Model.

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

All Case Studies highlighted the need for tool development for digital preservation. There are significant cost reductions possible, in ingest and metadata creation.

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

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

The LIFE project team encourage you to read the full project documentation on the accompanying CD.

The project documentation can also be found at **http://www.ucl.ac.uk/ls/lifeproject/**

LIFE Project : Documentation and deliverables on the CD

LIFE Project Summary
A short Report providing an overview of the Project's results and findings.

Research Review
A detailed literature review that describes the background to the Project, and the selection and development of the methodology and lifecycle approach.

LIFE Project Final Report
The Report describes the Project's approach, methodology and findings in developing lifecycle techniques to identify and cost the preservation of digital materials. The Report includes:

- A full description of the LIFE methodology and Lifecycle Model
- 3 detailed Case Studies which apply the cost model to the areas of Voluntary Legal Deposit at the British Library, Web Archiving at the British Library, and e-journals at UCL.
- An in-depth description and discussion of the Generic LIFE Preservation Model which is a tool for estimating the cost of performing preservation activities
- Findings and conclusions from the Project

Preservation documentation
Spreadsheets providing the detailed cost estimations for preservation activity for both the VDEP and Web Archiving Case Studies.