LIBER QUARTERLY, ISSN 1435-5205

Igitur. Utrecht Publishing & Archiving Services

How much does it cost? The LIFE Project - Costing Models for Digital Curation and Preservation

by RICHARD DAVIES, PAUL AYRIS, RORY MCLEOD, HELEN SHENTON, PAUL WHEATLEY[1]

INTRODUCTION

Digital preservation is concerned with the long-term safekeeping of electronic resources. How can we be confident of their permanence, if we do not know the cost of preservation? The LIFE (Lifecycle Information for E-Literature) Project has made a major step forward in understanding the long-term costs in this complex area. The LIFE Project has developed a methodology to model the digital lifecycle and to calculate the costs of preserving digital information for the next 5, 10 or 100 years. National and higher education (HE) libraries can now apply this process and plan effectively for the preservation of their digital collections.

Based on previous work undertaken on the lifecycles of paper-based materials, the LIFE Project created a lifecycle model and applied it to real-life digital collections across a diverse subject range. Three case studies examined the everyday operations, processes and costs involved in their respective activities. The results were then used to calculate the direct costs for each element of the digital lifecycle. The Project has made major advances in costing preservation activities, as well as making detailed costs of real digital preservation activities available. The second phase of LIFE (LIFE^2), which recently started, aims to refine the lifecycle methodology and to add a greater range and breadth to the project with additional exemplar case studies.

LIFECYCLE COLLECTION MANAGEMENT

In November 2005 a comprehensive review of existing lifecycle models and digital preservation 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 or HE setting. This is a brief synopsis of the full 96 page review by Watson (Watson, 2005). The review introduced 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 a range of LCC work from the construction industry to the waste management industry, in order to find an appropriate methodology.

However, it was within the library sector LCC work that the greatest synergy was found. This work was closely aligned with the work that was started in 1988 by Andy Stephens (Stephens, 1988). He developed a formula for calculating the total cost of keeping an item in a library throughout its lifecycle. Stephens' work is significant as it is the first attempt found which takes a library-based approach to the lifecycle management of assets. Although quite obviously developed for the paper world there is a strong correlation between the stages of analogue and digital asset management.

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

This work was continued by Helen Shenton where a specific focus on the aspects of preservation costs throughout the lifecycle was included (Shenton, 2003). 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 LIFE METHODOLOGY

The LIFE model is shown in Figure 1. The lifecycle has been broken down into six key elements: Acquisition (Aq), Ingest (I), Metadata (M), Access (Ac), Storage (S) and Preservation (P). L is the complete lifecycle cost over time (T). Each of these six elements can be further broken down into sub-elements, listed in Figure 2.

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

Figure 1: The LIFE Model

Acquisition	Ingest	Metadata	Access	Storage	Preservation
Selection (Aq1)	Quality Assurance (I1)	Characteri-sation M1)	Reference Linking (Ac1)	Bit-stream Storage Costs (S1)	Technology Watch (P1)
IPR (Aq2)	Deposit (I2)	Descriptive (M2)	User Support (Ac2)		Preservation Tool Cost (P2)
Licensing (Aq3)	Holdings Update (I3)	Administrative (M3)	Access Mechanism (Ac3)		Preservation Metadata (P3)
Ordering & Invoicing (Aq4)					Preservation Action (P4)
Obtaining (Aq5)					Quality Assurance (P5)
Check-in (Aq6)					

Figure 2: Breakdown of elements in the LIFE Model

Costs can be incurred at different stages of the lifecycle, and can occur just once or a number of times at different frequencies. LIFE proposes the calculation of costs for a single title or entity over a specific time period. There are a range of costs specified in the case studies, which include:

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

LIFE GENERIC PRESERVATION MODEL

It became clear during the early stages of the project that the Preservation element of the model would need much greater development than the other elements, due to the lack of work done in the areas of digital preservation costing. With this in mind, the project produced the Generic LIFE Preservation Model. There were four key objectives for this area of the project:

- Making the first major step in defining and estimating the lifecycle cost of digital preservation activities;
- Proposing a model for comment by the wider preservation community;
- Providing some rough cost estimates for the P element in the Lifecycle Model;
- Attempting to identify the scale of preservation costs.

The key elements of preservation activities were identified and the factors which contributed to their costs were then 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 developed as part of the model. The model is summarised graphically in Figure 3. The preservation cost for a particular file format from time=0 to time=t consists of both a regular Technology Watch cost and less frequent spikes of preservation activity, when action is required to ensure continued access to the format. By examining the issues highlighted, institutions can start to reduce the spikes of cost, as well as the frequency of the preservation actions.

Performing a preservation action would include activities such as:

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

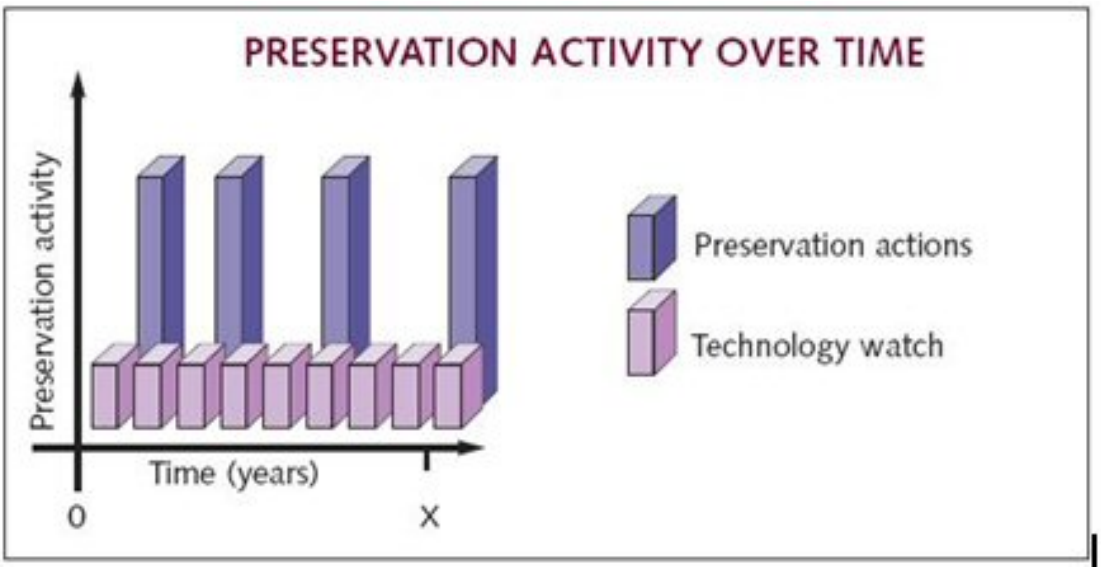

Figure 3: The Generic LIFE Preservation Model

The Generic LIFE Preservation Model is the first detailed attempt to identify and predict preservation costs. Use of the model provides no more than estimated costs with a degree of inaccuracy. It is hoped that further revision in phase 2 of LIFE will significantly improve the accuracy of the model. The revision exercise will incorporate input from economics review activities, review from a digital preservation expert group and the examination of more real life data. As is, the model provides a useful examination of component costs and a guide as to the serious economic commitments required for long-term preservation.

CASE STUDIES AND FINDINGS

Three exemplar case studies were chosen to test the LIFE methodology rigorously. They were: Web Archiving, E-Journals and Voluntarily Deposited Electronic Publications (VDEP) at the British Library (BL). The case studies have proven to be highly effective in highlighting both the types of issues that can be encountered in a digital collection, and the ways in which a lifecycle methodology can be utilised to capture and apply a cost to these problems. This combination of real data, and a framework within which to apply this information, has simplified many of the perceived areas of complexity within the digital preservation arena. Each of the case studies demonstrated that the LIFE Project Lifecycle Model proved more than suitable for calculating and comparing the costs of each activity. There were a number of practical and strategic findings for each of the case studies. A few key examples are highlighted here.

Web Archiving

The Web Archiving case study considered the costs of the BL's web archiving activities. Currently the BL is leading a collaboration with five other institutions as part of the UK Web Archiving Consortium (UKWAC). The aim of the UKWAC is to develop a test bed for archiving UK websites, and selectively to collect and preserve a cross-section of culturally significant web sites.

For the Web Archiving 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. The current Web Archiving activities are in their infancy in terms of scale, and also in terms of the capture of content. Collection and recording of metadata, the execution of characterisation of the content for the purposes of preservation, and the capture of the context of the selected sites are key areas for development. The costs of these operations need to be investigated.

Greater efficiencies, and the introduction of more automated processes, will reduce Web Archiving costs considerably, but unavoidable manual effort is likely to leave the costs of Ingest at a relatively high level for the medium term. The likely introduction of legal deposit legislation covering web materials will dramatically cut the cost of the IPR portion of the acquisition costs.

E-Journals

The e-journals case study was based at UCL Library Services. UCL acquires single titles, NESLi2 packages and non-NESLi2 packages which are reviewed annually.[2] UCL undertook two case studies in e-journals management: lifecycle collection management for the Public Library of Science corpus (PLoS) and for material from Blackwells.

UCL Library Services found that different elements of the Lifecycle Model fell under the spotlight when it analysed its own workflows and processes. UCL is geared towards giving access to e-journal literature, and to answering enquiries about the resulting access. The emphasis is not on Ingest, Storage or Preservation. However, it was possible for UCL to calculate, using activity-based costing, the total cost of making e-journals available to users. It was noted, however, that for most HE libraries, activity-based costing is not yet embedded in the workflow of the organisation.

UCL, as a research-led institution, has as its objective the acquisition of and access to, e-journal content for its staff and students. At the time of the case study, 8668 e-journal titles were logged in a UCL Access database. 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. Should university libraries each be responsible for digitally preserving the e-journals which they themselves acquire, or is it more cost effective to see this as an activity best performed for the sector by a trusted third party?

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 Preservation Model, VDEP preservation costs are projected to go down over time, not up, for this collection. There are, as yet, no obsolete file formats within VDEP and indeed LIFE struggled to find any formats at risk in any of its three case studies. Both Ingest and Metadata processes are currently very manual and in their present form incur a high proportion of the lifecycle cost.

Large-scale investment at the Ingest point to automate metadata would vastly reduce processing costs. Standard Metadata schema development is crucial for a digital repository. A national/international standard must be developed. The LIFE study clearly shows that Architecture standardisation is essential.

CONCLUSIONS

The process of identifying the different elements of a digital object's lifecycle, and then costing those elements, provided a very useful insight and approach to the challenges of digital preservation, beyond the obvious outputs of costing data useful for strategic planning. Performing a lifecycle-based costing exercise may provide some negative outcomes, particularly if sensitive cost or activity data is revealed to the outside world. LIFE has been somewhat courageous in exposing this kind of internal information and hopes that the benefits of this approach will be seen to outweigh the possible negative aspects.

The three case studies, which are vastly different in both content and workflow, have as expected returned three very different outcomes. However the variations in cost and workflow have been successfully captured within the lifecycle and preservation models. The VDEP's costs are strongly weighted in the areas of metadata and storage. This contrasts with the high acquisition and access costs for e-journals and Web Archiving's preservation costs. However, the LIFE model is able to capture all of these distinct trends and can therefore be used to capture a snapshot of any digital collection at any point in time.

Furthermore, the study was not able to reach a conclusion as to the relative cost-effectiveness of a national/copyright deposit library vis-à-vis an HE library taking on digital preservation as a national responsibility. As with paper archiving, the identification of just one body as having responsibility for digital archiving is a risk, as there is only one single point of failure. It should be noted, however, that the BL is further ahead than HE in terms of adopting and costing a lifecycle approach to the long-term management of digital assets.

All exemplars picked up on the fact that tool development for digital preservation is a high priority. There are significant costs to be saved, in areas such as Ingest and Metadata, if the correct tools are able to be developed for all aspects of the lifecycle of digital collections.

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

Both the lifecycle and the preservation models show that the LIFE work marks a significant step forward in the field of digital curation and preservation, and forms a unique piece of work. However, throughout the project, the team identified a number of areas where the LIFE work can be expanded in its second phase - LIFE^2.

FUTURE DEVELOPMENT – LIFE^2

The first phase of the LIFE Project ended in April 2006 with an international conference at the BL. It became clear, however, that further work was necessary. Both the LIFE methodology and the Generic Preservation Model would clearly benefit from further testing and revision with a wider range of case studies.

The second phase of LIFE (LIFE[2]) began in February 2007 after the team successfully bid for JISC funding for a further eighteen months. LIFE[2] will expand the work done in the first phase with four additional case studies – two institutional repository case studies as well as one for primary data, and a case study based on digitisation as surrogacy. These last two exemplars are a critical expansion of the methodology, as they include material that is not born digital (as was the case with the case studies in the original LIFE Project). LIFE[2] also includes a rigorous analysis of the validity of both models by an economics consultant.

REFERENCES

Mcleod, R., P. Ayris and P. Wheatley: *LIFE Project Report*. 2006. Available online from: www.life.ac.uk and http://eprints.ucl.ac.uk/archive/00001854/.

Shenton, H.: "Life cycle collection management". *LIBER Quarterly*, 13(2003), 254-272. http://liber.library.uu.nl/publish/articles/000033/article.pdf.

Stephens, A.: "The application of life cycle costing in libraries". *British Journal of Academic Librarianship*, 3(1988), 82-88.

Stephens, A.: "The application of life cycle costing in libraries: A case study based on acquisition and retention of library materials in the British Library". *IFLA Journal*, 20(1994), 130-140.

Watson, J.: "The LIFE Project research review: mapping the landscape, riding a life cycle." 2005. http://eprints.ucl.ac.uk/archive/00001856/01/review.pdf

WEBSITES REFERRED TO IN THE TEXT

British Library. http://www.bl.uk

LIFE - Life Cycle Information for E-Literature. http://www.life.ac.uk.

UCL - University College London Library Services. http://www.ucl.ac.uk/Library

UKWAC - UK Web Archiving Consortium. http://www.webarchive.org.uk

NOTES

[1] The article is written with contributions from Paul Ayris, Director of Library Services, University College London, Rory McLeod, Digital Preservation Manager,The British Library, Helen Shenton, Head of Collection Care,The British Library, and Paul Wheatley, Digital Preservation Manager, The British Library

[2] The initials NESLI stand for National Electronic Site Licence.