Costing the Digital Preservation Lifecycle More Effectively

Paul Wheatley

The British Library

Boston Spa, Wetherby, West Yorkshire, LS23 7BQ, United Kingdom

paul.wheatley@bl.uk

Abstract

Having confidence in the permanence of a digital resource requires a deep understanding of the preservation activities that will need to be performed throughout its lifetime and an ability to plan and resource for those activities. The LIFE (Lifecycle Information For E-Literature) and LIFE[2] Projects have advanced understanding of the short and long-term costs in this complex area, facilitating better planning, comparison and evaluation of digital lifecycles.

The LIFE Project created a digital lifecycle model based on previous work undertaken on the lifecycles of paper-based materials. It applied the model to real-life collections, modelling their lifecycles and studying their constituent processes. The LIFE[2] Project has reviewed and refined the costing model and associated tools, making it easier for organizations to study, cost and compare their digital lifecycles in a useful way. New Case Studies provided useful practical experience of the application of these costing tools and brought the LIFE approach full circle by investigating the comparison of complex digital and analogue lifecycles. The Case Studies were able to elicit useful results, although digital preservation lifecycle costing remains a complex and involved process.

The LIFE Project

The LIFE Project was funded by JISC to explore the costing of digital preservation activities using a lifecycle approach. The project ran for 12 months, ending in April 2006. It was a collaboration between The British Library (BL) and University College London (UCL).

Background and Research Review

The Project began with a comprehensive review of existing lifecycle models and digital preservation costing activities (Watson 2005). The concept of lifecycle costing, which is used within many industries as a cost management or product development tool is concerned with all stages of a product's or process's lifecycle from inception to retirement. The review looked at applications of the lifecycle costing approach in several industries including construction and waste management, in order to identify, assess and potentially reuse an appropriate methodology.

It was within the Library sector that the greatest synergy and potential for adaptation to the digital problem area was found. A model for estimating the total cost of keeping a print item in a library throughout its lifecycle provided a useful starting point (Stephens 1988). Although developed for the paper world, there were interesting parallels between the stages of analogue and digital asset management that would subsequently prove useful. The original model was later extended to cover preservation costs (Shenton 2003). The lifecycle stages start with selection, acquisitions processing, cataloguing and press-marking and continue through to preservation, conservation, storage, retrieval and the de-accession of duplicates. Three key "life stages" were selected as useful reference points at which to calculate costs. Year 1 provided an indication of initial costs following the significant selection and acquisition stages. Year 10 represented a review point and possible technological change or surrogacy. Year 100 was chosen as the symbolic "long-term" point, useful for forecasting downstream costs. Building on the foundations of this primarily print-focused lifecycle approach, LIFE developed a costing model for digital materials.

The LIFE Model

The LIFE Model v1.0 (Ayris, McLeod and Wheatley 2006) provided a content independent view of the digital lifecycle, breaking it down into Stages and Elements (see Figure 1). Each LIFE Stage represents a high-level process within a lifecycle that groups related lifecycle functions that typically occur or recur at the same point in time. These related functions are termed LIFE Elements. The LIFE model provided a common structure to which specific lifecycles could be mapped, enabling costing, analysis and comparison in a concise, readable and consistent manner.

The LIFE Methodology

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 utilising the

Acquisition	Ingest	Metadata	Access	Storage	Preservation
Selection	Quality Assurance	Characterization	Reference Linking	Bit-stream Storage Costs	Technology Watch
IPR	Deposit	Descriptive	User Support		Preservation Tool Cost
Licensing	Holdings Update	Administrative	Access Mechanism		Preservation Metadata
Ordering & Invoicing					Preservation Action
Obtaining					Quality Assurance
Check-in					

Figure 1: the LIFE Model v1.0, showing the breakdown of Stages (across the top) and Elements (down the page)

number of objects that would be processed. Staff costs were captured using studies of the involved personnel and the time they spent on different tasks. Costs were simply projected over time based on present day value, without consideration for inflation. LIFE calculated costs for 1, 5, 10 and 20 years.

The LIFE Case Studies

Three case studies were chosen for the application and evaluation of the LIFE Model and Methodology. They were:

- Web Archiving at the British Library
- Voluntarily Deposited Electronic Publications (VDEP) at the British Library
- E-Journals at UCL

The resulting lifecycle costs and the full workings of how these costs were calculated can be found on the LIFE website (www.life.ac.uk).

The Generic Preservation Model

The Case Studies considered by the first phase of LIFE did not contain activities addressing the preservation of content, such as technology watch, preservation planning or migration. With no preservation processes to observe and cost, an alternative strategy had to be pursued. Attention was focused on the development of a model to estimate the long-term preservation costs. The work of

Oltmans and Kol (2005) provided a useful starting point on which to build a more detailed model. Desk research and various team review and evaluation work led to the creation of the Generic Preservation Model (GPM). The GPM takes as an input a basic collection profile and provides as output estimates of the costs of preserving that collection for a certain period of time.

The LIFE[2] Project

While the LIFE Project was felt to have made significant progress in this difficult problem area, the project team felt that there was still much to do in advancing our ability to accurately assess, cost and compare digital lifecycles. Although the first phase of the project had devised a useful approach and had provided some indicative costs and analysis in case study form, a more thorough test, review and strengthening of this approach was necessary.

The LIFE team successfully applied for funding for a second phase of the project (LIFE[2]), which began in March 2007 and ran for 18 months. The British Library and UCL again implemented the project but added a number of Associate Partners to develop new Case Studies.

Review and Further Application Of The LIFE Approach

The Project started by initiating an independent assessment of the economic validity of the LIFE approach to lifecycle costing which was undertaken by Professor Bo-Christer

Creation or Purchase		Acquisition	Ingest	Bit-stream Preservation	Content Preservation	Access
....		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				

Figure 2: the LIFE Model v2.0

Björk from Hanken, the Swedish School of Economics and Business Administration (Björk 2007). The report largely validated the approach taken by the LIFE team and provided a number of recommendations to help steer the second phase of the project in the right direction.

The LIFE Model and Methodology was then reviewed and updated by the project team, using the independent assessment, as well as feedback gathered from the wider digital preservation community, as a foundation for this work. This resulted in version 1.1 of the LIFE Model (Wheatley, et al. 2007).

The revised LIFE tools were applied to new LIFE Case Studies, two of which were conducted at Associate Partner sites:

- SHERPA DP, which examined the lifecycle costs of a preservation service
- SHERPA-LEAP, which studied lifecycle costs at the institutional repositories of Goldsmiths at the University of London, Royal Holloway at the University of London, and UCL (University College London)
- British Library Newspapers, which studied and compared both analogue and digital lifecycles at this National Library

A fourth Case Study that had planned to examine the costs of primary data curation was not completed due to staffing issues at the Associate Partner site.

Lessons learnt from the experiences of the Case Studies were fed back into the LIFE approach resulting in a further release of the LIFE Model as version 2.0. Full details of the Case Studies and their findings can be found in the LIFE[2] Project Final Report (Ayris, et al. 2008) and key aspects of the LIFE approach that have enhanced our ability to cost digital lifecycles more effectively are discussed below.

LIFE[2] Developments

LIFE[2] invested considerable effort in developing the LIFE Model, Methodology and associated tools in order to improve the accuracy and consistency of the costing process, to simplify the work involved and to ensure that the results of lifecycle costing activities could be usefully applied.

An assessment of digital preservation costing objectives was undertaken, with the aim of identifying where the application of lifecycle costing data would be useful, and thus informing the development of the tools used to capture that data. Key objectives included:

- Identification of selective costs, such as repository running costs
- The cost of adding a new content stream lifecycle to an existing repository
- Evaluating the efficiency of an existing content stream lifecycle

- Assessing the impact of a new tool or a process change within an existing content stream lifecycle
- Comparison of similar lifecycles at different organisations
- Comparison of analogue and digital preservation

This assessment provided useful guidance in the development of the scope of lifecycle costing, which is addressed in more detail in the LIFE[2] Final Report.

The LIFE Model was revised following collation of a range of feedback on the LIFE[1] work. The LIFE team also liaised closely with the digital preservation costing team at the Danish Royal Library, State University Library and State Archives who provided invaluable comment and contribution as the Model was developed. The resulting release of the LIFE Model v2.0 provided a more detailed and more clearly defined picture of the digital lifecycle. Significant changes included clearer terminology, new lifecycle elements, particularly in Bit-stream Preservation, and more detailed definitions. As well as further description at the Stage and Element level, suggested Sub-element descriptions were included. These low-level lifecycle functions provide an indication of the scope and level of detail that would be useful to capture in a costing exercise, and most were found to be applicable for the lifecycles encountered in the Case Studies.

Conclusions

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 is not always a straightforward process. The revised and more detailed Model has reduced ambiguity. The Sub-element detail provides clearer guidance on the process of matching particular lifecycle processes to the LIFE Elements. 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. Despite these improvements, the addition of further detail to the Methodology would be desirable. This might include a tighter definition of the key processes and more guidance for users on the key costing procedures. While reviewing the LIFE Model, the team envisaged a categorization of cost types (e.g. capital, staff, development) and a more formal approach for capturing, costing and projecting them. Unfortunately, there was insufficient time to implement this. With the benefit of hindsight, it is clear that a more rigorous Methodology would have been useful, and should ideally have been prioritised over other developments.

Capturing the costs of lifecycles that are no longer actively ingesting digital objects proved to be problematic.

Although this was considered as a possible risk while planning the Case Studies in LIFE[2], it was not expected that this approach would be as time consuming as it turned out to be. Further difficulties were experienced in capturing a sufficient level of detail (with clear working) at the Associate Partner sites. As noted above, a more detailed methodology would have helped, but in contrast with the LIFE[1] Case Studies, it was clear that costing activities are far easier to lead within the managing organisation's own realm of responsibility. Far more effort was required to implement the LIFE[2] Case Studies than was expected, and this placed a considerable strain on project resources.

The complex nature of the lifecycles examined in the British Library Newspaper Case Study provided a thorough test of the LIFE approach for comparing and contrasting analogue and digital costs. The Case Study was able to elicit results that allowed some useful comparisons to be made, but the complexity involved highlighted that these analogue and digital mappings were very much in their infancy. The LIFE team is keen to further develop and explore our ability to compare and contrast analogue and digital lifecycle costs with the ultimate aim of informing the difficult digital versus analogue collection management decisions looming on the horizon.

Considerable progress has been made in costing the digital preservation lifecycle, despite the relatively small effort that has so far been directed at this complex and multi-faceted problem area. Since the start of LIFE[1], other new developments have emerged. A consortium of Danish organizations, including the National Library, State and University Library and the State Archives, are developing the LIFE Model for cross-institutional comparison of their digital preservation activities. JISC also funded a study into the costs of data curation, which utilized elements of the LIFE work (Beagrie, Chruszcz and Lavoie 2008). Despite these advances, digital preservation costing remains in its infancy and our current tools can provide us with indicative but not accurate digital preservation costs.

Moving forward our ability to cost the digital preservation lifecycle will require further investment in costing tools and costing 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 ingested. In order to support these developments a greater volume of raw cost data will be required to inform and test new cost models. Organisations undertaking digital preservation activities are therefore encouraged to record costs as they proceed and where possible make their figures available to the wider community.

Looking ahead to LIFE[3]

A third phase of the LIFE work is currently under consideration. Initial proposals include a focus on development of an integrated toolset to both streamline the process of costing an existing digital lifecycle and estimate the cost of implementing a new lifecycle. The predictive tool would take as an input a simple profile of a new digital collection or content stream and a profile of the preserving organisation. The tool would then automatically process these profiles and estimate the costs for each lifecycle stage for a required timescale.

Acknowledgements

References

Watson, J. 2005 *The LIFE project research review: mapping the landscape, riding a life cycle*. London, UK. http://eprints.ucl.ac.uk/1856/1/review.pdf

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

Shenton, H. 2003. *Life Cycle Collection Management* LIBER Quarterly 13, 254-272. http://liber.library.uu.nl/publish/articles/000033/article.pdf

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

Oltmans, E., Kol, N. 2005. *A Comparison Between Migration and Emulation in Terms of Costs*. RLG Diginews 9, http://worldcat.org/arcviewer/1/OCC/2007/08/08/0000070519/viewer/file959.html#article0

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

Wheatley, P., Ayris, P., Davies, R., Mcleod, R. and Shenton, H. 2007. *The LIFE Model v1.1*. LIFE Project, London, UK. http://eprints.ucl.ac.uk/4831/

Ayris, Davies, Mcleod, Miao, Shenton, Wheatley, 2008. *The LIFE[2] Final Project Report*. LIFE2 Project, London, UK. http://www.life.ac.uk/2/documentation.shtml

Beagrie, N., Chruszcz, J., and Lavoie, B. 2008. *Keeping research data safe: a cost model and guidance for UK Universities*. JISC, London, UK. http://www.ndk.cz/dokumenty/dlouhodoba-ochrana/keeping-research-data-safe-a-cost-model-and-guidance-for-uk-universities