LIFE: costing the digital preservation lifecycle
iPRES 2007

Paul Wheatley, paul.wheatley@bl.uk
The British Library, Boston Spa, Wetherby, LS23 7BQ, UK.

Paul Ayris, ucylpay@ucl.ac.uk
UCL, Gower Street, London, WC1E 6BT, UK

Richard Davies, richard.davies@bl.uk
The British Library, St Pancras, London, NW1 2DB, UK.

Rory McLeod, rory.mcleod@bl.uk
The British Library, St Pancras, London, NW1 2DB, UK.

Helen Shenton, helen.shenton@bl.uk
The British Library, St Pancras, London, NW1 2DB, UK.

EXECUTIVE SUMMARY
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) Project[1] has 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 results were then used to estimate the costs of each element of the digital lifecycle. Organisations can now apply this process, enabling evaluation and refinement of their existing lifecycles and facilitating more effective planning for the preservation of newly acquired content.

Phase 2 of the LIFE Project began in February 2007. It is evaluating and refining the models and methodology developed in the first phase of the project and developing lifecycle costings for a range of further case studies.

LIFECYCLE COLLECTION MANAGEMENT
In November 2005 a comprehensive review of existing lifecycle models and digital preservation was undertaken (Watson, 2005). This was conducted in order to find a useable cost model that could be applied to the management of digital collections within a Library or Higher Education setting.

The review introduced the concept of lifecycle costing, 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 applications of the lifecycle costing approach in several industries including construction and waste management, in order to find and potentially reuse an appropriate methodology.

---

[1] LIFE and LIFE[2] are collaborative projects led by UCL and the British Library, and funded by the Joint Information Systems Committee (JISC).

It was within the Library sector lifecycle costing work that the greatest synergy and potential for adaptation to the digital problem area, was found. Stephens (1998) developed a model for estimating the total cost of keeping a print item in a library throughout its lifecycle. Although developed for the paper world, there are interesting parallels between the stages of analogue and digital asset management.

Stephens returns to this work in 1994 and identified costs for monograph and serials holdings of the national collection at the British Library. This work was continued by Helen Shenton (2003) who extended the original model to cover preservation costs across the lifecycle. 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 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.

Figure 1: The LIFE Model.

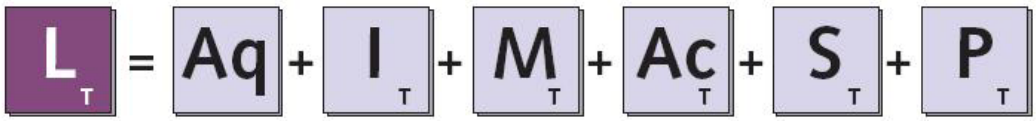


Figure 2: Breakdown of lifecycle elements in the model.

Acquisition	Ingest	Metadata	Access	Storage	Preservation
Selection	Quality Assurance	Characterisation	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					

Acquisition is driven by the all important Selection process where new acquisitions are chosen. The right to access and preserve is sought during the IPR and Licensing processes, followed by administration activities involving Ordering and Invoicing. The content is physically transported to the preserving organisation in the Obtaining phase, and is then briefly verified to ensure the expected content has been received in a Check-in process.

Ingest begins with a detailed Quality Assurance process which assesses whether the content is what it purports to be and is of a sufficient level of quality. The content is committed to the repository during the Deposit phase. Holdings Update adds details of the new content in appropriate holdings records.

Key metadata is created and recorded throughout the Metadata stage, which begins with an automated Characterisation process, where file formats are identified and content is validated. Processes for recording Descriptive and Administrative Metadata follow.

Access begins with the creation of reference information which facilitates the finding of the content, which is termed Reference Linking. The system for providing access to the content and associated processes is considered in the Access Mechanism element and provision for User Support is also covered.

Bit-stream preservation is addressed in the Storage stage. These processes were not broken down any further in the LIFE1 Model, but will be expanded on in the second phase of the LIFE project.

The Preservation stage begins with monitoring of the technical environment and provision of obsolescence alerts. Tool support for performing preservation actions is the next significant step. Preservation Metadata, including representation information is then created as appropriate. Preservation Actions are performed to ensure continued usability of content. These are then verified in a Quality Assurance phase.

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 and the numbers of objects that would be processed. Staff costs were captured using studies of the involved personnel and the time they spent of 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.

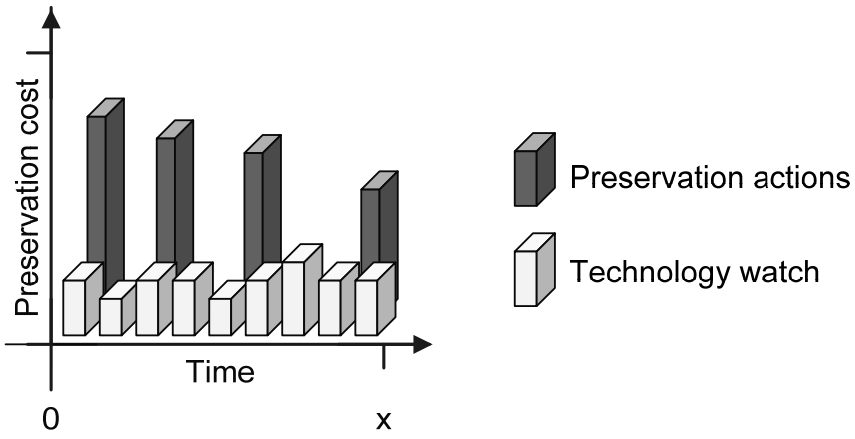
Costs can be incurred at each stage of the lifecycle. Costs may be incurred just once, may accrue over time, or recur on a regular or irregular basis. The case studies highlighted this cross section of cost types including one-off costs in the first year for content Selection, costs that accrue over time such as Storage, and recurring costs for Preservation. The methodology enables the estimation of costs for a single title, item or instance over a given time period.

THE GENERIC LIFE PRESERVATION MODEL

The case studies considered by the project 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 and, it was hoped, more accurate model.

The key preservation activities were identified and the factors which contributed to their costs were then modelled. Costs fell into two main categories: annually recurring technology watch and planning activities and less frequent but more intensive preservation actions. Figure 3 provides an illustrated example of the occurrence of these key processes and costs over time.

Figure 3: The occurrence of preservation costs over time.



Simple equations were developed to model these preservation processes and costs. Significant inputs to the model were included as editable constants that could be set by users of the model as appropriate to their situation (for example, software developer costs). Trends were estimated and modelled in areas such as the availability of software tools over time or the economies of scale of batch processing content. Figure 3 hints at two of these trends. It shows the frequency of the execution of preservation actions reducing over time as the introduction of new file formats becomes less frequent and existing formats become more stable and are standardised. The cost of preservation actions reduces over time as the availability of preservation tools increases due to increasing external investment in digital preservation infrastructure.

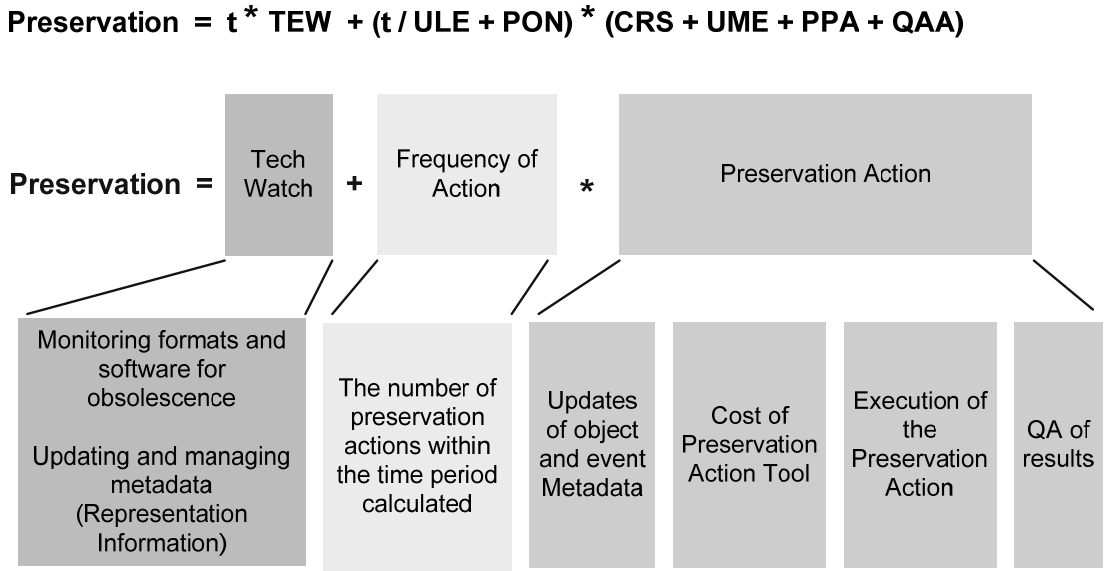
The model maintains a largely neutral view of the differing preservation strategies (normalisation, format migration, migration on request, emulation) that might be utilised to preserve digital content. This generic view was pursued in an attempt to avoid any contentious arguments for and against such opinion polarising approaches as emulation and migration. The timing of preservation action is however critical when it comes to cost calculation over a specific period of time. An easily editable input was therefore provided to model the percentage of content normalised (migrated on ingest) as opposed to being migrated or emulated at or around the point of obsolescence. This was nominally set at 40%.

The resulting model enabled the estimation of costs for the preservation of a number of digital objects of a particular format. Given basic information on the content profile of a particular collection (extracted with the aid of a tool such as DROID[2]) the costs of preservation over time could then be estimated. This in turn could be used to calculate an average cost for the preservation stage of the lifecycle.

[2] DROID is a file format identification tool, that works alongside the PRONOM technical registry: http://en.wikipedia.org/wiki/PRONOM_technical_registry

Figure 4 shows the Generic LIFE Preservation Model with a graphical expansion and explanation of its key components.

Figure 4: The Generic LIFE Preservation Model



Expansion of terms:
TEW – TEchnology Watch
ULE – Unaided Life Expectancy
PON – Proportion Of Normalisation
CRS – Cost of new Rendering Solution
UME – Update Metadata
PPA – Performing Preservation Action
QAA – Quality AssurAnce
A further detailed expansion of the terms used, including suggested values for the model inputs, can be found in the LIFE Project Report (2006).

CASE STUDIES AND FINDINGS

Three case studies were chosen for the application and evaluation of the LIFE Model and Methodology. They were: Web Archiving, and Voluntarily Deposited Electronic Publications (VDEP) at the British Library, and 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. A brief summary of the case studies and some key findings based on the costing work is given 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) to selectively collect and archive a cross-section of culturally significant web sites.

The current Web Archiving activities are in their infancy in terms of scale and the method of content capture. 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 will 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 experienced by UKWAC, by removing the need to seek archiving permissions from the content owner.

E-Journals

The e-journals case study was based at the Library Services of UCL, a research-led Higher Education institution with a focus on the provision of e-journal content for its staff and students. UCL acquires single titles as well as large packages of e-journals (both NESLi2[3] packages and non-NESLi2 packages) which are reviewed annually. At the time of the case study, UCL had 8668 e-journal titles licensed for use within the institution. UCL elected to examine two subsets of e-journals which included titles from the Public Library of Science corpus (PLoS) and titles 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. Ingest and storage are provided by the publishers themselves. Responsibilities for preservation are unclear. However, it was possible for UCL to calculate the total cost of making e-journals available to users, by careful study and costing of the activities involved. It was noted, however, that for most HE libraries, activity-based costing is not yet embedded in the workflow of the organisation.

VDEP

Voluntarily Deposited Electronic Publications (VDEP) housed at the BL provided the final case study. The VDEP content has been acquired under voluntary legal deposit legislation for digital materials. A wide range of content has been collected, totalling some 230000 objects at the time the case study was conducted. Average lifecycle costs were calculated, although the wide range of content types, physical size, and frequency of issues made it difficult to summarise costs across the 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. Investment at the Ingest point to automate metadata creation and capture would vastly reduce processing costs.

All case studies clearly identified tool development as a high priority. Preservation tools and infrastructure as well as Ingest tools and Metadata capture facilities demonstrate the most significant potential for automation and cost saving. Targeted investment in these areas will be essential to bring the lifecycle costs down.

---

[3] NESLi2 is the UK's national initiative for the licensing of electronic journals on behalf of the higher and further education and research communities. http://www.nesli2.ac.uk/

CONCLUSIONS

The LIFE Model was able to capture and identify key trends in the project case studies and demonstrated potential for further use in a number of roles:

- Improved assessment of the financial commitment an organization is making when acquiring or creating new digital materials.
- More effective planning for future preservation activities.
- Comparison of digital lifecycles across an organisation or between different types of organisation.
- Evaluation and optimisation of existing digital lifecycles.
- Generation of guidance to funding bodies, such as JISC, to address the aspects of the digital lifecycle which would most benefit from an investment in tool development and automation.

Adopting institutions are now beginning to explore the benefits of taking a lifecycle approach to digital preservation planning and costing. The LIFE Project partners, UCL and the BL, have already begun to embed the LIFE approach in their everyday digital preservation activities. A number of Danish institutions have also adopted the LIFE approach, including the Royal and State Libraries, the National Archives and the National Film Institute. The Royal Library is, at the time of writing, embarking on a new project to establish lifecycle costing, based on the LIFE Model, across Danish cultural heritage institutions.

The Generic LIFE Preservation Model is the first detailed attempt to identify and predict preservation costs. It is hoped that further revision and refinement will significantly improve the accuracy of the model. 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.

THE LIFE PROJECT, PHASE 2

The second phase of LIFE ($LIFE^2$) began in February 2007 and has started with a thorough review of the LIFE Models and Methodology by an independent economics expert. The results of the review, in addition to feedback gathered from the project team and adopting institutions are guiding a revision of the Model and a more detailed realisation of the LIFE Methodology.

$LIFE^2$ 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 model, as they include material that is not born digital (as was the case with the case studies in the original LIFE Project).

The ultimate vision for the LIFE work is to provide automated predictive costing tools for each element of the lifecycle. These tools would take, as an input, assumptions about technology development and the profile of the content an organisation might be acquiring or creating. Automated tools would use this information to provide estimates for the cost and effort required to preserve the content over a particular period.

REFERENCES

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

Shenton, H. (2003) 'Life Cycle Collection Management' in *LIBER Quarterly* vol. 13, pp. 254-272. Available online at: http://liber.library.uu.nl/publish/articles/000033/article.pdf.

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

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

Watson, J. (2005) The LIFE Project research review: Mapping the landscape, riding a life cycle. Available online from: www.life.ac.uk and at http://eprints.ucl.ac.uk/archive/00001856/.

Oltmans, E., Kol, N,. (2005) "A Comparison Between Migration and Emulation in Terms of Costs" in *RLG Diginews*, Volume 9, Number 2 ISSN 1093-5371. Available online at: http://www.rlg.org/en/page.php?Page_ID=20571#article0

WEBSITES REFERRED TO IN THE TEXT

LIFE Project Website at http://www.life.ac.uk