iPRES 2009

THE SIXTH INTERNATIONAL CONFERENCE ON THE PRESERVATION OF DIGITAL OBJECTS

Proceedings

October 5-6, 2009
Mission Bay Conference Center
San Francisco, California

California Digital Library

LIFE^3: Predicting Long Term Digital Preservation Costs

Paul Wheatley and Brian Hole
The British Library
Boston Spa, Wetherby, West Yorkshire, LS23 7BQ, United Kingdom
paul.wheatley@bl.uk

Abstract

As we develop our ability to preserve digital collections through techniques such as migration and emulation, the decision process of what action to take and when to take it becomes increasingly complex. Cost is a crucial factor to consider but the financial implications of preservation planning decisions are not typically well understood. At a strategic level, there are also significant challenges to contend with as the world moves rapidly to a world of both non-digital and digital information provision. What is the appropriate size and make up of an organisation's preservation department?

A new phase of the LIFE Project is aiming to improve our understanding of the financial aspect of these questions, ensuring preservation risk is minimised and preservation activity can be conducted within the boundaries of our financial constraints.

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 approach supported comparison and analysis of digital preservation activity across the complete lifecycle. LIFE^3 is now beginning to look to the future with the development of a predictive costing model that will support more effective decision making and planning for digital preservation.

The LIFE Project, Phases 1 and 2

The British Library (BL) and University College London (UCL) were co-funded by JISC in the first two phases of the LIFE Project. Beginning in 2005 with LIFE^1 and continuing in 2007 with LIFE^2, the Projects explored a lifecycle approach to mapping out and costing digital preservation activities.

LIFE^1 focused on developing a usable approach to lifecycle costing and drew on a number of case studies that examined the costs of digital preservation activity at UCL and the BL. LIFE^2 evaluated and refined this approach through external review and the application of LIFE techniques to a wider range of lifecycles at different organisations.

Background and Research Review

The LIFE work 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 and methodology for digital materials.

The LIFE Model

The LIFE Model v2.1 (Ayris, Davies, McLeod, Miao, Shenton, Wheatley 2008) was developed to provide a content neutral view of the digital lifecycle from the perspective of the preserving organisation. The lifecycle was broken down into six high level LIFE Stages representing the functions associated with preservation and access. These Stages were in turn divided into LIFE Elements which represented specific lifecycle functions (see figure 1) The Element level captured lifecycle processes at a level of granularity that was sufficiently high to be relevant across a range of different content and organisational types while still providing useful detail from an analytical stand point. Maintaining a standardised

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 1: the LIFE Model v2.1

and generic view across different organisations or content types facilitated comparative analysis of different lifecycles while presenting costing information in a concise, readable and consistent manner.

Version 2.1 of the LIFE Model introduced a further layer of decomposition, with LIFE Sub-elements. Sub-elements provided additional description by suggesting likely components of element level processes. These sub-processes are not considered to be standardised across different lifecycles but instead facilitate understanding and assist with the identification of likely lifecycle processes.

The LIFE Methodology

The LIFE Methodology was developed to provide guidance on studying an existing lifecycle and recording the component costs of processes at each lifecycle stage. An initial process of establishing the scope and time frame of the case study is followed by the identification of relevant processes and staff and initial interviews to inform the drafting of a graphical lifecycle workflow. This workflow captures lifecycle processes in terms of organisation and content. These are then mapped to the LIFE Model and reviewed and refined with those staff responsible for conducting the actual lifecycle work. Cost capture and analysis can then be conducted. Typical costs might include those of equipment, setup and ongoing staff. An appropriate method of capturing these key costs is chosen and applied. Capital costs are averaged across their expected lifetime based on the number of objects to be processed. Staff costs are captured using studies of the involved personnel and the time spent on lifecycle relevant tasks. Costs are then projected over time based on present day value.

Case Studies in Phases 1 and 2 of LIFE

A range of case studies were chosen for the application and evaluation of the LIFE Model and Methodology across the first two phases of the LIFE Project. They were:

- Web Archiving at the British Library
- Voluntarily Deposited Electronic Publications (VDEP) at the British Library
- E-Journals at UCL
- SHERPA DP, which examined the lifecycle costs of a centralised 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)
- Newspapers at the British Library, which studied and compared both analogue and digital lifecycles

A fourth Case Study that had been planned to examine the costs of primary data curation was not completed due to staffing issues at the Associate Partner site. 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 preservation watch, preservation planning or migration. With no Content 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 expert review and evaluation work led to the creation of the Generic Preservation Model (GPM). The GPM provided the ability to estimate Content Preservation costs based on a basic content profile and a range of configurable inputs. The initial GPM model developed in LIFE[1] was refined in the second phase of the project, and then reviewed at the beginning of 2009 by a cross organisational expert group. Recommendations from this meeting will be addressed in further work throughout LIFE3. Collaboration with a number of Danish memory organisations, including the Royal Library, will continue through the third phase of LIFE. This work is currently focusing on developing a model for estimating migration costs (Bøgvad Kejser 2009).

The LIFE[3] Project

Aims

The LIFE[3] Project, which began in August 2009, is moving the focus of the LIFE work from retrospective costing and post-event analysis to predictive costing and a supporting role in enhancing planning and decision making activities.

As memory organisations move closer to providing comprehensive support for digital materials and research projects generate ever greater amounts of digital output, it is becoming critical to have a clear picture of the necessary levels of resource required to support preservation. Even with a dedicated Digital Preservation Team, the British Library has over twenty times more effort dedicated to non-digital preservation than it has to digital preservation. This ratio is expected to change over the next few years, but it remains unclear how far it will need to move.

Ongoing digital preservation costs beyond the first year of implementation are still relatively poorly understood which makes even short to medium term resource planning a challenge. This issue is brought further into focus by the context of the current move from a predominantly nondigital to a more closely balanced hybrid world.

LIFE[3] aims to improve our ability to anticipate the resourcing needs of future digital preservation activity, guiding decision making over whether or when to acquire, how or when to preserve, and how much resource needs to be put in place over the longer term.

Estimative Costing Tool

LIFE[3] will develop an estimative costing tool that will generate costs for a particular period of preservation activity given details of the organisational context, the current technological environment and a description of the content in question. The organisational inputs will be captured in an organisational profile which provides details of policy, legal constrains and current status of existing preservation activity. Configurable inputs will enable the current state of the art, for example in hardware storage capability and cost, to be captured and maintained in an up to date fashion. A content profile will gather key details of the digital material whose preservation will be costed, including details of file formats and the number and size of the digital objects. These input profiles will be processed by a series of mathematical models developed from the GPM and refined and extended through the use of case study data and an expert review process, which will generate estimated costs for each stage and element of the LIFE Model.

The LIFE tool will then be integrated with a new costing module of the DRAMBORA risk assessment tool. A stand alone version will also be made available, and where possible care will be taken to use a data schema suitable for integrating this with other preservation tools as well.

Process and Current Status

The estimative models are currently being developed using an iterative process allowing creation, testing, and refinement. Model development is well underway with a current focus on the creation of a new Bit-Stream Preservation model and revisions of the existing GPM Model for the Content Preservation Stage of the lifecycle. Existing and related current work is being exploited where appropriate. Development of the software tool that incorporates the costing models will begin early in 2010. Sources of information for the input profiles to the estimative costing tool are currently under consideration, with the aim of making use of existing toolsets where possible, and automating information gathering for the user. It is hoped that the organisational profile will be populated at least in part using data from DRAMBORA. Developments on the Planets Project (Planets 2009) in capturing an organisation's preservation policy in a machine interpretable form are also likely to elicit useful input information. The necessary content profile will be generated by the Planets Profiling Service.

Template Approach

As well as developing a low level modelling approach that will analyse detailed inputs and provide specific estimated costs on output, a template approach will be explored to generate quick estimates with a minimum of effort. This will make the tool more useful to a wider range of users, from researchers needing a quick cost estimate for a funding proposal, to institutions wanting to work out a detailed long term strategy. The user of the costing tool will be able to choose from templates representing broad organisational and content profiles which will automatically populate the detailed inputs to the model. As well as utilising templates developed by LIFE, it will be possible to generate custom templates tailored to particular organisational settings. LIFE^3 will experiment with this approach while trialling the iterations of the predictive models with the aim of assessing its usefulness.

Applying LIFE at the British Library

As the third phase of the LIFE Project moves forward from research to realisation of the LIFE approach with the development of a costing tool, the Digital Preservation Team at the British Library is beginning to exploit the work of LIFE. At the strategic level there are questions on the size and composition of the preservation department of the future and on the status of digital as an effective preservation medium. At the operational level there are the specific questions of what to preserve, when to preserve it and how much resource will be required to do the job while minimising preservation risk to an acceptable level. The LIFE developments are now beginning to play a significant role in answering some of these complex questions.

The Lifecycle Approach

The lifecycle approach can usefully be applied to structure, and where necessary, join up activity to manage digital materials. Experiences at the BL have shown how crucial it can be to consider digital preservation requirements at each stage in the lifecycle of a digital object, ensuring preservation is pre-emptive and efficient. Reactive or remedial preservation activities have been shown to be costly, as was experienced with activity to stabilise and preserve digitised masters that had not been monitored or managed closely over a period of years.

The LIFE work builds on well establish theory on the proactive approach of lifecycle management (Beagrie and Jones, 2001), providing structure and ensuring visibility of preservation activity across the lifecycle. Developing a unified approach to preservation, whether digital or nondigital, remains a key goal for the BL Digital Preservation Team and the Collection Care department within which it is partially based. While there will remain specialists in both digital and non-digital fields, many key preservation roles will not be specific to the nature of the content being preserved. The lifecycle approach will underpin this unified preservation strategy as the BL continues to redefine its approach to preservation.

A key aim from the very beginning of the LIFE work was to facilitate a better understanding of a key collection management decision facing the BL Collection Care Department: the appropriateness of digital as a preservation medium for non digital materials. The key factors that the BL is aiming achieve satisfactory control over are the cost of the approach and the preservation risk it is subject to. The experiences of costing an array of digital and non-digital preservation activities in a comparable way has dramatically increased the BL's understanding of this critical balance. As a result of this progress, a move from microfilm to digital surrogacy has therefore become increasingly possible.

Collection Management Decision Making

Making key collection management decisions without an appreciation for the medium or long term implications on preservation and resourcing can leave a ticking time bomb for later in the lifecycle. This is particularly crucial where acquisition by purchase or digitisation to create a new digital collection is facilitated by external funding. In this case the focus is typically on the short term issues of acquisition and access but it is vital to consider the commitment to activity later in the lifecycle which is typically not supported financially by the external funder.

The LIFE case studies, described above, have begun to provide the evidence to support organisational change by demonstrating the considerable potential for efficiency savings. By investing a little more up front, a substantial amount of resource can typically be saved over the medium and longer terms. This might include activities such as putting in place effective validation and ensuring appropriate technical standards are followed. As well as increasing the efficiency of lifecycle activity, the effect is also to reduce preservation risk.

A greater understanding of preservation costs can ensure more effective resource planning as well as facilitating smarter acquisition decisions. If a new collection is to be acquired, an estimative lifecycle costing tool provides the ability to plan for preservation effort beyond the very short term, instead of managing this work reactively.

Preservation Planning

When pre-emptive or remedial action must be taken to ensure the longevity of digital information, typically where a file format or access mechanism has become obsolete, a preservation planning process is necessary to enable the selection of an appropriate treatment. The BL

has begun to apply a formalised preservation planning process on its digital collections using the Planets Preservation Planning tool, Plato (Becker 2007). Gathering the data to inform this decision making process remains a challenge but is critical in achieving a satisfactory preservation outcome. In particular, a clear indication of the relative costs of the preservation options under consideration is vital.

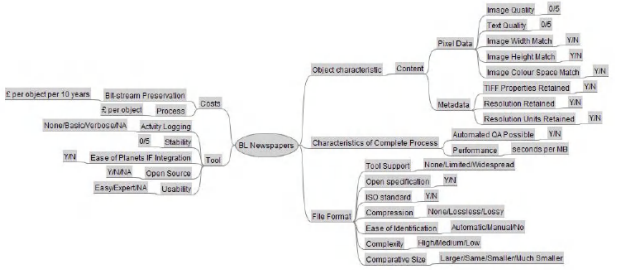

Figure 2: preservation planning requirements for a digitised newspaper collection at the BL

Figure 2 shows a mind map of requirements for a preservation planning process for a digitised newspaper collection at the BL. The preservation plan resulting from these requirements considered the costs of implementing a chosen preservation option (labelled as "process costs") as well as longer term bit-stream preservation costs. This information was generated by embryonic developments from LIFE[3]. The facility was not available to consider the impact of costs across the whole lifecycle, but it is hoped this will be possible by utilising later versions of the LIFE[3] costing tool. While it is expected that other stages of the lifecycle will not be as significant to the longer term total cost as the Content and Bit-stream Preservation Stages, factors such as the resulting size of migrated files will impact substantially on cost Elements such as the re-ingest of content to a digital repository.

Digital preservation activities at an organisation like the British Library are typically conducted on a significant scale (operating on an 80 Terabyte collection, in the example above). This brings considerations of cost to the fore, and places increased emphasis on the need to effectively predict the cost implications of preservation decisions. The need to balance cost with preservation risk is expected to be the one of the key challenges faced over the next few years.

Acknowledgements

Thanks go to the LIFE Project Team and the many other staff of The British Library and UCL, without whom the results of the Project would not have been realised. The current phase of the project has been made possible by funding from the Joint Information Systems Committee and the Research Information Network and is supported by LIBER.

References

Ayris, P., Davies, R., McLeod, R., Miao, R., Shenton, H., Wheatley, P. 2008. *The LIFE[2] Final Project Report*. LIFE2 Project, London, UK. http://eprints.ucl.ac.uk/11758/

Beagrie, N., Jones, M., 2001, *Preservation Management of Digital Materials: The Handbook*, http://www.dpconline.org/docs/handbook/DPCHandbook.pdf

Becker, C., Strodl, S., Neumayer, R., Rauber, A., Nicchiarelli Bettelli, E., Kaiser, M., 2007, *Long-Term Preservation of Electronic Theses and Dissertations: A Case Study in Preservation Planning*, Ninth Russian National Research Conference, RCDL'07

Bøgvad Kejser, U., 2009, "Cost Model for Digital Preservation: Cost of Digital Migration," *iPRES 2009 Proceedings*

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

Planets Project website, 2009, http://www.planetsproject.eu

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

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

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