LIFE^3: A PREDICTIVE COSTING TOOL FOR DIGITAL COLLECTIONS

Brian Hole
The British Library
St Pancras, 96 Euston Road,
London, NW1 2DB,
United Kingdom

Li Lin
The British Library
St Pancras, 96 Euston Road,
London, NW1 2DB,
United Kingdom

Patrick McCann
HATII
George Service House
11 University Gardens
University of Glasgow,
Glasgow ,G12 8QH, Scotland

Paul Wheatley
The British Library
Boston Spa, Wetherby, West
Yorkshire, LS23 7BQ,
United Kingdom

ABSTRACT

Predicting the costs of long-term digital preservation is a crucial yet complex task for even the largest repositories and institutions. For smaller projects and individual researchers faced with preservation requirements, the problem is even more overwhelming, as they lack the accumulated experience of the former. Yet being able to estimate future preservation costs is vital to answering a range of important questions for each. The LIFE (Life Cycle Information for E-Literature) project, which has just completed its third phase, helps institutions and researchers address these concerns, reducing the financial and preservation risks, and allowing decision makers to assess a range of options in order to achieve effective preservation while operating within financial restraints. The project is a collaboration between University College London (UCL), The British Library and the Humanities Advanced Technology and Information Institute (HATII) at the University of Glasgow. Funding has been supplied in the UK by the Joint Information Systems Committee (JISC) and the Research Information Network (RIN).

1. INTRODUCTION

Life Cycle Collection Management has been described as "a very complex subject with many practical, financial and strategic interdependencies" [1]. The LIFE model and tool make an important contribution to approaching this subject by providing costing estimates for the lifecycle of digital collections, and consequently allowing for the exploration of the practical and strategic dimensions as well. Stakeholders with an interest in this area include libraries, archives and museums, as well as research and Higher Education (HE) institutions along with the individual researchers within them. As part of their mandate to provide access to their collections for the long term, the greatest concerns that they have involve collection management, technology strategy, human resource management, and central to all of these, budgeting and funding.

The following are examples drawn from recent literature of where costing information could be used to address questions in each of these areas. With a continual influx of material, libraries are constantly forced to make difficult decisions regarding the balance of their collections, such as whether to retain less used physical items due to pressure on storage space [2]. Knowing the true cost of digitising items is important when comparing this to other options such as continued physical storage, disposal and reassignment of space for other purposes [3].

In terms of technology strategy, digital repositories are becoming extremely important as central components of institutions' technology infrastructures [4]. Knowing the relative costs is essential in choosing the correct repository and preservation system, where the future financial consequences of mistakes can be serious [5].

Institutions are often unsure as to their human resource requirements as the digital proportions of their collections increase. Should the related work be done in-house, outsourced, in collaboration with other organisations [6], or by re-training existing staff [7]?

Determining the true cost of a digitisation project and being able to justify it is critical, as most institutions have to seek external funding for such work [8]. Not taking medium and long-term preservation factors into consideration can create a "ticking time bomb" [9], which requires additional, unplanned funding to diffuse at a later date. Organisations need to understand that funding for digital preservation needs to be provided on an ongoing rather than temporary basis, and how to incorporate planning for this into their budgets [10]. Grant applications need to include sound design, a detailed management plan (including a commitment to preservation), a complete and realistic budget, details of all required human resources, and plans for effective sustainability [11].

Institutions require an understanding of the costs of the entire digital lifecycle in all of the above situations in order to ensure sustainability and preservation [12], especially as the preservation actions involved at each stage may not be initially obvious to them [13]. The LIFE model and tool provide an accessible and practical way of determining these costs, in order that these critical decisions can be made with greater confidence.

2. BACKGROUND

The LIFE project has so far run over a total of three and a half years, spread over three phases. The first phase ran from 2005 to 2006. This established that a lifecycle approach to costing digital collections was applicable and useful, and developed a methodology for doing so. It tested this approach by applying it to real life

collections in a number of case studies, including Voluntarily Deposited Electronic Publications (VDEP) and web archiving at the British Library, and the e-journals repository at UCL. It also developed a model for estimating the preservation costs of a digital objects lifecycle [14].

This was followed by phase two in 2007 and 2008, which included further validation of the model, economic assessment of the LIFE approach and further testing and evidence generation via additional case studies. These included the SHERPA-LEAP institutional repositories, SHERPA-DP digital preservation services, and the British Library Newspapers digitisation project. Feedback from the LIFE[2] final conference indicated considerable demand for a predictive costing tool to aid in planning digital preservation [15].

3. LIFE[3]

The third phase which ran from 2009 to 2010 and has just completed, has delivered a web-based predictive costing tool that significantly improves the ability of organizations to plan and manage the preservation of digital content. This tool is based upon a refined version of the LIFE model produced in phase two (see figure 1), following collection of additional case study and survey data. This has enabled the model to cover a wider range of preservation scenarios, including sound, web and e-journal archiving, in addition to print.

Creation or Purchase	Acquisition	Ingest	Bit-stream Preserv.	Content Preserv.	Access
Creation	Selection	Quality Assurance	Repository Admin.	Preserv. Watch	Access Provision
	Submission Agreement	Metadata	Storage Provision	Preserv. Planning	Access Control
	IPR & Licensing	Deposit	Refresh	Preserv. Action	User Support
	Ordering & Invoicing	Holdings Update	Backup	Re-ingest	
	Obtaining	Reference Linking	Inspection	Disposal	
	Check-in				

Fig. 1 The LIFE model

A survey of digital preservation repositories was carried out in order to better understand their storage requirements and costs, with these being correlated to the size and purpose of each system. Aside from the number of mirror sites employed, the survey looked at the combination of storage technologies used for access as well as backup, the cost and expected lifetime of the hardware, and also as other factors such as support, infrastructure and electricity costs.

3.1. Model Development

This data was then collected and built into a financial model, using Excel and Visual Basic. The Excel workbook includes a basic input sheet, the output sheet which displays the calculated costs for all the stages, six data refinement sheets that allow the user to modify estimations used within each model stage, and six model sheets that contain the financial models used for calculating costs throughout the lifecycle. The Visual Basic code involves a number of subroutines that are linked with macros to perform functions such as filling and clearing input cells within the workbook.

While the model is designed to produce accurate estimates due to a thorough understanding of the preservation lifecycle and associated variables, it was felt that it should also be able to provide quicker estimates for the purpose of comparison, where many options under consideration can be quickly discounted. A template approach was followed to allow the user to select from content and organisation categories into which their particular project falls. The model is then populated with default data calculated from the mean values of case studies that also fall into those categories.

A user thus has to enter data into only five fields on the basic input sheet in order to receive an in initial cost estimate. These are simply the time frame of the project, the original media type of the material to be preserved (print, website, sound, research material, or other) the source (purchased, donated or to be created through digitsation or harvesting), the number of items to be processed in each year of the project, and the size of institution involved. In the case of digitisation, they are also asked for the quality required. This information is used to pre-populate the model with data averaged from relevant case studies where it is available, and the user is immediately presented with a cost estimate on the output page. They are able to drill down and change the default values at each stage of the life cycle in order to achieve a more precise result using the refinement sheets, or they can simply reset the model and try a different configuration (see figure 2). All figures on the output page are rounded to two significant figures in order to underline the fact that they are indicative estimates only, and users are made aware of the fact that case study data is illustrative rather than absolute. Initial numbers are likely to be higher than expected because it is assumed that all stages of the lifecycle are being carried out, often defaulting to more conservative scenarios.

The first thing the user is likely to do is adjust the model to use the infrastructure and staff costs specific to their institution. The ‘Refine Organisational Profile’ sheet will contain default data based on the size of institution the user has selected on the input sheet, but unless the sheet has been previously modified by someone in the same organisation these are unlikely to be accurate.

Users can choose the number of storage sites to be modelled, along with the storage technology and its cost for each one, as well as for backup. Technologies included in the model are spinning disk, enterprise tape, flash storage, and pay per use (e.g. cloud storage). One of the highest security factors for preservation is diversity of storage methods and vendors [16], so the ability to experiment with different scenarios and supplier costs here is very useful. Staff costs based on annual, daily or hourly rates should also be entered here for the five project roles used in the model, from Senior Manager to Operational Staff. These rates are used throughout the model wherever staff costs are calculated. For UK HE institutions, users can also enter the indirect and estate figures for each role to ensure proper calculation of Full Economic Costs (FEC). Staff costs are then adjusted for inflation across time.

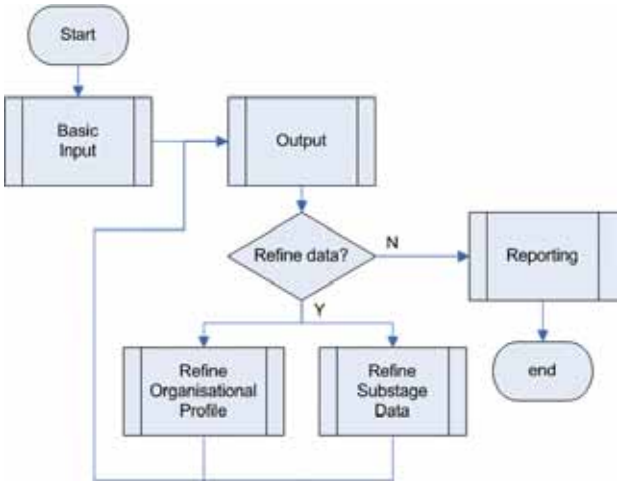

Fig. 2 Typical workflow

The 'Creation or Purchase' stage calculates costs based on the source chosen by the user on the input sheet. For purchased items, the total purchase cost is derived by summing-up the purchase cost of all years (purchase cost per item x number of items per year). For donated items, the cost at this stage is simply zero. For digitised items, the user is presented with 23 digitisation cost elements across three columns to capture small, medium and large projects, and the associated case study derived default data. Elements are either based on labour costs (e.g. days of work for a project manager to shape the project) or cost per item digitised (e.g. deshelving and capture). Users should check each one of these figures, correcting them where necessary or setting them to zero when a task is not part of the project under consideration. This challenges institutions to justify non-inclusion of best practice tasks such as QA and metadata capture.

For the 'Acquisition' and 'Ingest' stages, the user is able to adjust the default data for 35 cost elements, based on hours, days, or percentage of time spent on each by staff members of a certain role. As ingest is an area where the KRDS2 project noted that there are potential savings to be made by many projects [17], users should use this section to experiment and try to find cost savings.

The 'Bitstream Preservation' stage allows the user to edit the costs for repository administration, refreshment, backup and administration. In addition to this, the costing factors for each type of storage technology can be changed, including lifetime, cost per MB, rate of cost deflation (applied throughout time) and electricity costs. As the latter cost is especially significant for enterprise systems [16] users should pay attention that this is correct for their region or institution. It is important to note that the technologies we employ today are not permanent solutions however [12], and that we really cannot predict what will be available in 20 years [21], so all model predictions beyond this point should really be accepted with great caution.

It was noted at the end of the LIFE^2 phase that the 'Content Preservation' stage still required development [15], and this has now been simplified and reworked, taking into consideration the work of the Danish national library and archives [18]. Each content type is assigned a heterogeneity level describing the number of different file formats involved of high (e.g. websites) or low (e.g. print), and a complexity level regarding these files of high (e.g. MS Word or PDF documents) or low (e.g. tiff files). The combination of each of these factors is then used to determine the cost of any content migrated. Users are given three migration strategies to choose from, these being 'do nothing', 'migrate on ingest', and 'migrate periodically'. In the case of 'do nothing', users can also enter a cost for emulation. This is the chosen strategy for the KB in Holland for example, betting on the stability of emulation in the long term [16]. The Welcome Library on the other hand count on the fact that by accepting only a limited range of formats thought to be stable, the 'do nothing' option will work without emulation [19]. For 'migrate on ingest', the cost of migration is calculated for each year of the project based on the number of items selected. In the case of 'migrate periodically' the user can determine the percentage of items to be migrated and the number of years between migrations. It is recommended that users challenge their assumptions and experiment with these options, as the costs within this section of the model can be significant depending on the options chosen. It has been noted that institutions should not count on the falling cost of storage, as a growing number of items due to migration can easily offset these gains [13], while the operational costs of some preservation strategies may actually exceed the perceived value of a collection [16]. Rusbridge has also cautioned that the assumption that file formats become obsolete rapidly and that interventions should thus be made on a frequent basis is likely to be incorrect in many cases where until recently it was an accepted truth [20].

Finally, the 'Access' stage provides default estimates for the costs of creating, maintaining and managing an access system, based on both direct costs and staff effort. Users are also able to determine whether some costs will recur periodically due to replacement or refreshing of the system.

The LIFE3 model has been exposed to members of the digital preservation community during its development, and has received very positive feedback, in particular due to its immediate usability.

3.2. Web Tool Development

In conjunction with HATII, a web-based tool incorporating the financial model has been produced. The aim of the tool is to make the LIFE model both easily accessible and easy to operate for all levels and backgrounds of users. As an example of this, when using the tool in comparison to the spreadsheet, only the data that is directly relevant to the user at any point in time is displayed. Once the user has drilled down into the data and edited it to the point that they feel it is representative of their project, they are able to produce a full report of the predicted cost and all of the factors that have been involved in calculating it. This can not only be used to demonstrate the thoroughness of the prediction, but is a useful checklist for users to make sure that they have in fact taken all required tasks into account.

The application has been developed using the open-source Symfony (http://symfony-project.org) object-oriented PHP framework on top of a MySQL database. PHP and MySQL are well-established open-source technologies in which the developers at HATII have plenty of experience. The use of an MVC framework allowed the development to proceed more rapidly and provided a standard, well-documented structure.

In order to ensure ease of sustainability for the tool in future, it was required that the economic model employed by the application be able to be edited by an administrator without the need for a developer. This meant that as much of the logic of the model as possible had to be contained within the database, the structure of which was kept as general as possible.

The following can be considered a description of the model in the context of the application in the broadest terms possible. A preservation project takes place over a number of years. It is classified in a number of ways (category, source, organisation type etc.) and a number of items are processed each year. The model can be thought of as a set of properties that can be used to describe a project. The value of a property of a given project can be drawn from a case study, entered by the user or calculated from the values of other properties of the project. The ways the project is classified determines which properties apply and how their values are determined.

The basic entities can be seen in this description: project, project year, property, value, classification and category. But much of the power of the model comes from the way in which property values are derived from each other through calculations. To provide the necessary configurability therefore, those calculations also needed to be stored in the database. Simple arithmetic formulae using the sum, product, difference and quotient operators can be easily evaluated when they are described using postfix notation (http://scriptasylum.com/tutorials/infix_postfix/algorithms/postfix-evaluation/index.htm). The algorithm involves reading the expression from left to right, so the formulae are stored in the form of a linked list of components in the database. Each component is either an operand or an operator, and where it is an operand it contains a reference to the property whose value is to be used in the application. The postfix evaluation algorithm can hence be applied quite simply.

A challenge involved in this approach is that the performance of the application can be adversely affected by the need to retrieve not just data for calculation input but the calculations themselves from the database as they are evaluated. Also, any of the values supplied as operands to a calculation may have to be calculated themselves. Another issue is that many properties need to be assigned values for each year of a project, so the number of entities involved in the calculation of an estimate increases greatly as the length of the project increases. PHP's limitations when it comes to managing memory use when executing object-oriented code (specifically garbage collection of objects containing circular references) means that every opportunity needs to be taken to avoid creating objects in memory and to destroy them correctly once they are finished with.

Some specific aspects of the model have had to be handled differently to the standard calculation structure described above. The application of economic factors to costs that recur over each year of a project and costs that occur on a periodic basis are two examples. This logic has therefore had to be written into the application, though the recurrence period and the economic factors themselves remain customisable.

Generally, however, the approach to complications not catered for by the implementation of the model has been to increase the flexibility of the model rather than to implement specific solutions. For example, as it became apparent that additional types of classification were necessary, and that an administrator would need control over them (e.g. organisation size was been added to the model after development began), these were abstracted away from the project object, allowing the behaviour of the model to be tailored according to all of the possible combinations of classification applied to the project.

Most importantly, the tool has been designed to be easily maintainable by its hosting institution, without the need for further programming. All variables and formulas used in the model can be edited through a user administration interface. In this way the financial model can be modified to take account of new factors (for example a new task or additional hardware requirement) and any errors in formulas can be fixed.

4. FUTURE WORK

While LIFE has to date produced an extremely valuable resource for the digital preservation community, future work will ensure that this resource is widely available and of maximum use going forward. This will focus on making the LIFE tool widely available as a working and sustainable service with promotion, support and knowledgebase maintenance and enhancement. It will also make the service more applicable to a wider range and type of institutions globally, by internationalizing the financial model and extending the breadth and depth of the data.

To do this, LIFE will partner with the Open Planets Foundation (OPF), a new foundation with a global footprint that is dedicated to providing technology, advice and on-line complimentary services for the planning of digital preservation. OPF will provide hosting, promotion, support and maintenance, effectively taking LIFE from a functioning tool to a working, sustainable Service.

The Service will be further developed based upon controlled evaluation with selected HE/FE partner sites, and the accuracy of LIFE cost estimation will also be enhanced by establishing a process for collating and integrating new costing data in the LIFE knowledgebase. Finally, the LIFE Service will also be internationalized in order to improve its usability worldwide, with support for different currencies and a wider range of international data.

5. REFERENCES

[1] Shenton, H., "Life Cycle Collection Management", *Liber Quarterly* 13:254-272, 2003.

[2] Holt, G. E., "Economic Realities in Optimizing Library Materials Access", *The Bottom Line: Managing Library Finances,* 20(1):45-49, 2007.

[3] Robinson, C. K., "Library Space in the Digital Age: The Pressure is on", *The Bottom Line: Managing Library Finances,* 22(1):5-8, 2009.

[4] Jacobs, N., "Institutional Repositories in the UK: The JISC Approach", *Library Trends,* 57(2):124-141, 2008.

[5] Seadle, M., "The Digital Library in 100 Years: Damage Control", *Library Hi Tech,* 26(1):5-10, 2008.

[6] Middleton, K., "Collaborative Digitization Programs: A Multifaceted Approach to Sustainability", *Library Hi Tech,* 23(2):145-150, 2005.

[7] Shenton, H., "From talking to doing: Digital preservation at the British Library", http://dx.doi.org/10.1080/13614530009516807 (accessed July 11th 2010), *New Review of Academic Librarianship,* 6(1): 163-177, 2000.

[8] Eden, B., "Getting Started with Library Digitization Projects: Funding Your First (and Subsequent) Digital Projects", *The Bottom Line: Managing Library Finances,* 14(2):53-55, 2001.

[9] Wheatley, P. and Hole, B., "LIFE3: Predicting Long Term Digital Preservation Costs", http://www.life.ac.uk/3/docs/ipres2009v24.pdf (accessed 11th July 2010), paper presented at iPres 2009.

[10] Lavoie, B. and Dempsey, L. "Thirteen Ways of Looking at… Digital Preservation", *D-Lib Magazine,* 10(7/8), http://dlib.org/dlib/july04/lavoie/07lavoie.html (accessed July 5th 2010), 2004.

[11] Ray, J., "Digitization Grants and How to Get One: Advice from the Director, Office of Library Services, Institute of Museum and Library Services", *The Bottom Line: Managing Library Finances,* 14(2), 2001.

[12] Bradley, K., "Defining Digital Sustainability", *Library Trends,* 56(1):148-163, 2007.

[13] Chapman, S. "Counting the Costs of Digital Preservation: Is Repository Storage Affordable?", *Journal of Digital Information,* https://journals.tdl.org/jodi/article/viewPDFInterstitial/100/99 4(2): 1-15, 2004.

[14] McLeod, R. and Wheatley, P. and Ayris, P. *Lifecycle information for e-literature: full report from the LIFE project.* http://eprints.ucl.ac.uk/1854/ (accessed 11th July 2010), LIFE Project, London, UK, 2006.

[15] Ayris, P. and Davies, R. and McLeod, R. and Miao, R. and Shenton, H. and Wheatley, P. *The LIFE2 final project report.* http://eprints.ucl.ac.uk/11758/ (accessed 11th July 2010), LIFE Project, London, UK, 2008.

[16] Rosenthal, D. S. H., Robertson, T., Lipkis, T., Reich, V. and Morabito, S., "Requirements for Digital Preservation Systems: A Bottom-Up Approach", arXiv:cs/0509018v2 [cs.DL], 2005.

[17] Beagrie, N., Lavoie, B. and Wollard, M., *Keeping Research Data Safe 2,*

http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf (accessed 11th July 2010), Charles Beagrie Ltd., 2010.

[18] Kejser, U. B., Nielsen, A. B. and Thirifays, A., *The Cost of Digital Preservation: Project Report v. 1.0,* Danish National Archives and Royal Library, 2009.

[19] Thompson, D., "A Pragmatic Approach to Preferred File Formats for Acquisition", http://www.ariadne.ac.uk/issue63/thompson/ (Accessed July 11th 2010), *Ariadne* 63, 2010.

[20] Rusbridge, C. "Excuse Me… Some Digital Preservation Fallacies?", *Ariadne* 46, 2006.

[21] Steele, K., "The Fiscal Wonders of Technology", *The Bottom Line: Managing Library Finances,* 22(4):123-125, 2009.