Proceedings of the 11th International Conference on Manufacturing Research (ICMR2013), Cranfield University, UK, 19th – 20th September 2013, pp 133-138

MODELLING LONG TERM DIGITAL PRESERVATION COSTS: A SCIENTIFIC DATA CASE STUDY

Essam Shehab, Alice Lefort
Mohamed Badawy, Paul Baguley
Christopher Turner
Manufacturing and Materials Department
Cranfield University
Cranfield
Bedfordshire, MK43 0AL, UK
e.shehab@cranfield.ac.uk
m.badawy@cranfield.ac.uk

Michael Wilson
Esther Conway
The Science and Technology
Facilities Council (STFC)
Harwell
Didcot
Oxfordshire, OX11 0QX, UK
michael.wilson@stfc.ac.uk
esther.conway@stfc.ac.uk

ABSTRACT

In recent years there has been increasing UK Government pressure on publicly funded researchers to plan the preservation and ensure the accessibility of their data for the long term. A critical challenge in implementing a digital preservation strategy is the estimation of such a programme's cost. This paper presents a case study based on the cost estimation of preserving scientific data produced in the ISIS facility based at The Science and Technology Facilities Council (STFC) Rutherford Appleton Laboratory UK. The model for cost estimation for long term digital preservation is presented along with an outline of the development and validation activities undertaken as part of this project. The framework and methodology from this research provide an insight into the task of costing long term digital preservation processes, and can potentially be adapted to deliver benefits to other organisations.

Keywords: Digital Preservation, Curation, Cost Estimation, Whole Lifecycle Cost

1 INTRODUCTION

The long term preservation and accessibility of data is a pressing issue for publicly funded researchers. In light of this a question arises for the researcher and the funding agency: how is such preservation funded and managed in a sustainable way? It is only possible to address this question if there is a commonly understood basis on which the long-term preservation costs can be predicted.

In the case of scientific data there are several good reasons why having a digital preservation policy is essential. Data that is properly preserved can be used to validate analyses published of it and scientific conclusions drawn from it; data may be reused in meta-studies to discover effects which were not identifiable in one data set alone; it can be used to set parameters in, and test new theories, and it can be used to address issues which were not considered at the time of the original experimentation (for example, several data sets collected several hundred years ago have recently been used to model and predict climate change). Scientific data may be unique, unrepeatable and time dependent so it cannot be collected again; or the experiments, from which data are collected, may be expensive to perform.

The Science and Technology Facilities Council (STFC) is one of Europe's largest multidisciplinary research organisations. STFC is a UK government body that supports a national and international community of more than 10,000 scientists. ISIS is a world-leading centre for research in the physical and life sciences at the Rutherford Appleton Laboratory near Oxford in the UK. ISIS is effectively a large microscope, which uses neutrons and muons instead of light, to determine the properties of materials at the scale of atoms, without harming them. ISIS operates about 30 different instruments, at

any time, which provide different details of the structure of a material. Hundreds of experiments are performed annually at ISIS by visiting researchers from around the world, in diverse science areas including physics, chemistry, material engineering, earth sciences, biology and archaeology (STFC, 2012).

A critical challenge in implementing a digital preservation strategy is the estimation of such a programme's cost. This paper presents a case study based on the cost estimation of preserving scientific data produced in the ISIS facility and presents the methodology used for the costing and the selection of the chosen method of data preservation. Working with STFC the authors of this paper have been able to:

- Define the work breakdown structure (or work flow) for long-term digital preservation in ISIS
- Identify the cost drivers for the ISIS facility data preservation
- Provide estimates for five preservation scenarios developed by the STFC e-Science research team through three different activities (pre-archive, archive and access)

2 RELATED RESEARCH

Although cost modelling for the long term preservation of data from facilities such as ISIS has not been studied before, studies in several areas associated with cost modelling for long-term digital preservation are relevant. These studies can be split between the categories of long-term digital preservation and cost estimation. There have been many authors who have attempted to precisely define Long-Term Digital Preservation (LTDP) (Lee et al., 2002; Borghoff, 2006). Factor et al. (2009) offers the following definition of LTDP as the set of "processes, strategies and tools used to store and access digital data for long periods of time during which technologies, format, hardware, software and technical communities are very likely to change."

The "digital data" mentioned by Factor et al. (2009) includes both the contents of the digital document and its metadata, which is the information describing the digital document (Borghoff, 2006). LTDP also aims to ensure the usability, the authenticity, the discoverability and the accessibility of the data (Lee et al., 2002). The definition provided by Factor et al. (2009) implies that there are the risks involved in digital preservation. Breeding (2010) briefly describes these risks as the "vulnerabilities of digital content"; the risk of data corruption and inaccessibility.

Ensuring digital preservation is not only about copying digital information; information must also be accessible for future processing (Lee et al., 2002). However, the main challenge is to interpret the digital document format (Borghoff, 2006) by producing an identical effect to that achieved on the original system. To answer this challenge, Lee et al. (2002) and Borghoff (2006) classify the LTDP techniques into two different technical approaches:

- The preservation of the original technological environment suitable for restoring the document in its original format,
- The continual transformation into the newest format while preserving the original "look and feel"

Table 1: Preservation Techniques: Advantages and limitations

Techniques	Key advantages	Limitations
Technology preservation	Ideal for short-term preservation: the authenticity cannot be better	Space, cost and durability of the devices
Technology emulation	Does not imply to save ageing original hardware and software but still keep authenticity	Requires precise, detailed and complex-to-create specifications and increases the amount of data to be preserved
Information	Well-known and widely used by IT	Reduces authenticity and increases com-

migration	services (several methods and tools exist)	plexity (different components require different migration activities)
Encapsulation	More flexible and consistent with the emulation and migration techniques	The way of how the strategy may be practically implement is not clear

To standardise digital preservation methods and provide a set of common best practices, several organisations have published guidelines, including the Open Archival Information System (OAIS) Reference Model. The OAIS Reference Model considers the context of the archive, as illustrated in Figure 1: An archive aims to store documents, submitted by a producer and delivered to a consumer; these activities are supervised by a management service (CCSDS, 2002).

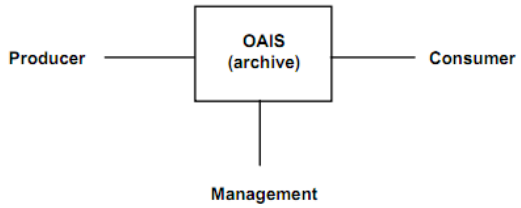

Figure 1: Environment model of an OAIS archive

The Life cycle Information For E-literature (LIFE) projects are a range of three consecutive projects based on existing long-term digital preservation cost modelling projects. They aim to deliver a cost model that can be used in the widest number of cases.

The Cost Model of Digital Preservation (CMDP) project is based on the OAIS Reference Model that provides a well-defined standardised breakdown of the relevant activities. It consists in the seven OAIS functional entities (Ingest, Data Management, Archival Storage, Access, Preservation Planning and Administration) as well as the three roles (Producer, Consumer and Management). CMDP project is based on the Activity-Based Costing technique.

3 RESEARCH METHODOLOGY

This paper aims to develop a cost model for the long-term digital preservation of data produced by the ISIS facility. In order to deliver this aim it has been necessary to employ a 3 phase methodology. The three main phases are: (i) understanding the context, (ii) data collection and analysis, (iii) model development and validation.

The first phase concentrated on examining the literature in long term preservation and cost estimation. The second phase concentrated on data collection and analysis in order to populate the cost model. This involved administering a questionnaire to more than 20 interviewees drawn from a range of roles within ISIS and STFC. The aim of the questionnaire was to establish the understanding of digital preservation held by the workforce of the two organisations and obtain the detailed responses required to accurately populate a cost model. Analysis of the data collected from the questionnaire was performed in the phase three. A cost model was then developed from the analysis and validated through two workshops held with experts drawn from industry.

4 PRESERVATION SCENARIOS FOR THE CASE STUDY

From the literature review (phase 1 in Figure 2) and by analysing the responses gained by the questionnaire (Phase 2 in Figure 2) it was clear that the encapsulation preservation technique was most appropriate for the scientific data at ISIS. A key feature of the encapsulation technique is that everything required to read a given data file must be archived with it. In the ISIS facility data is stored in the Nexus file format. The Nexus file is read by Mantid software, which can read, analyse and graphically visualise data inside the Nexus file.

4.1 Scenarios

In order to manage the risk of losing data (Conway et al., 2011) several preservation scenarios have been developed. The first preservation scenario, as shown in Figure 2, archives only the essential data. Although this is sufficient for short-term preservation while the current Mantid application is still used by the community, there is a high risk of the data becoming unreadable by future applications. The second preservation scenario removes such risks by, among other elements, emulating the original operating system compatible with the original data file and application as illustrated in Figure 3 (Conway and Lambert, 2011). Between these two extremes there are another three options.

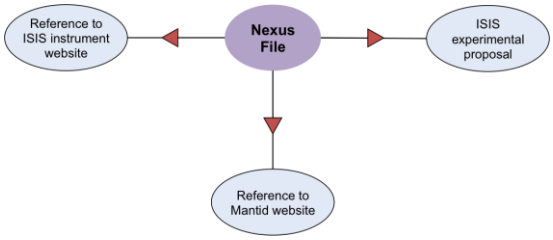

Figure 2: Scenario 1 – Basic Preservation

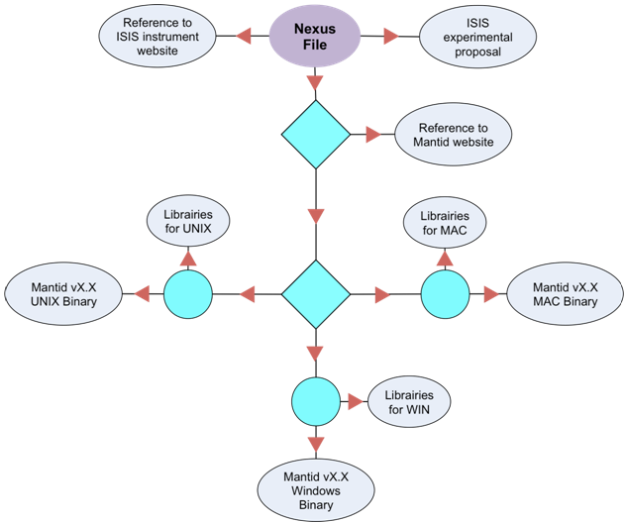

Figure 3: Scenario 2 – Extended Preservation

5 DEVELOPMENT OF THE COST MODEL

The development of the cost model was based on a combination of the analysis of the completed questionnaires and guidance from literature. The questionnaire was designed to be used in face-to-face meetings and to last from 30 minutes to one hour to fill, depending on the interviewee's field of expertise.

5.1 Cost Estimation Model

The aim of the cost model is to calculate the total preservation cost, at the end of Y-years preservation, when data has been accessed. This cost C(Y) is so represented by Equation (1).

$$C(Y) = C_{PRE-ARCHIVE}(Y) + C_{ARCHIVE}(Y) + C_{ACCESS}(Y) \quad (1)$$

Sixteen cost elements were defined and can be divided into four categories, and so are calculated in four different ways:

- Human costs (related to people activity) or non-human costs (not related to people activity) respectively represented by Equations (2) and (3), and
- Recurring and non-recurring costs (in this case, Recurrence equals 1 in. Equations (2) and (3).

$$Human\ Cost = \frac{Annual\ Human\ Cost \times N \times Time\ needed}{Recurrence\ of\ the\ activity} \quad (1)$$

Where:

Annually Human Cost is made of the salary overvalued by 65% for the overheads.

N is either the number of people needed or the number of times the activity has to be done in one recurrence.

$$Non-human\ Cost = \frac{Numbers\ needed \times Buying\ Price}{Life\ Span\ or\ Recurrence} \quad (2)$$

5.2 Overall Cost Model Structure

The cost model development for this project consists of several different modules. The overall cost model structure is represented in Figure 4. It describes the connections between the different modules to deliver the cost estimations to the user.

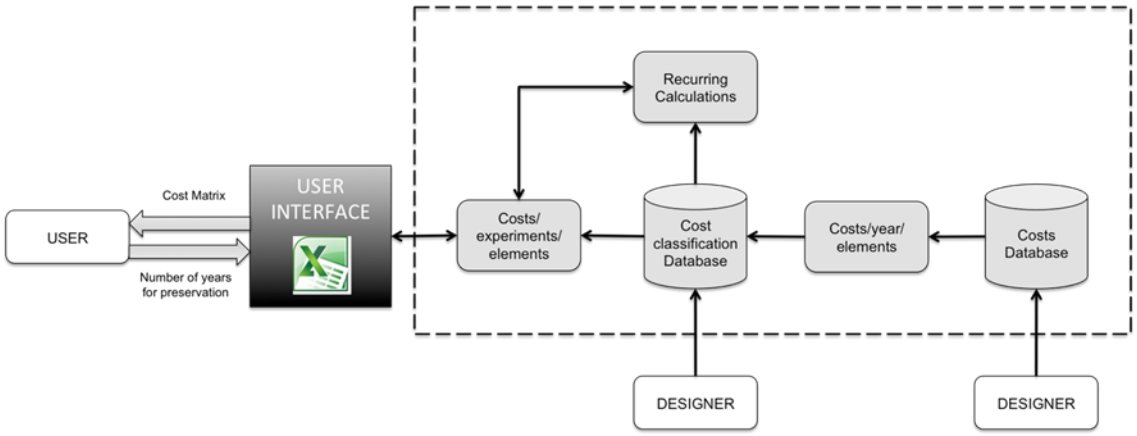

Figure 4: Cost Model Structure

6 VALIDATION

At present no ISIS scientific data long-term preservation cost estimation has been performed. Thus, the model was not validated through case study. However, two validation sessions were arranged with several experts in digital preservation and cost modelling. In the first validation session the developed cost model has been validated through a first expert judgment during a validation session. Four experts were invited to attend this session. The validation session lasted for around an hour. The general feedback provided on the cost model has been that it was meeting expectations of the experts. There is an appreciation of the level of detail in each cost element and the transparency of these costs. However, they noticed the lack of context around the scenarios and the assumptions inside the cost model: a user with a knowledge of what is digital preservation will have understood the phases (pre-archive, archive, access) but no one outside ISIS and e-Science will have understood the activities defined without previous knowledge of the project.

A second validation session was set up. More than 20 digital preservation experts from ISIS. The general feedback was that the cost model could provide a valuable insight into preservation costs at ISIS. An expert from the first validation session declared that: "The costs are following the trends I was expecting them to follow. Moreover, the range of costs seems to be realistic."

7 CONCLUSIONS

This research has delivered a model capable of estimating ISIS scientific data costs over the long-term. The model provides a matrix grouping of the three preservation phases costs (pre-archive, archive and access) through five preservation scenarios. It also compares these costs to the experiment rerun cost. In terms of cost modelling, long term digital preservation is a new area of research. This project will be used by STFC as a starting point for a larger project that they are partners in: Enabling kNowledge, Sustainability, Usability and Recovery for Economic Value (ENSURE), partly funded by the European Union. The ENSURE projects aims to provide new long-term digital preservation technologies. ENSURE will be based on three case studies from health care, clinical trials and finance and will focus on a number of issues, which have not been fully addressed by the literature up to this point.

ACKNOWLEDGMENTS

The authors would like to thank the staff of the Science and Technology Facilities Council and the partners of the ENSURE EU FP7 project for supporting this research. ENSURE project is supported by the Commission of European Community (contract number: 270000) under the ICT Programme. The authors acknowledge the European Commission for its support as well as the other partners in the consortium (www.ensure-fp7.eu).

REFERENCES

Borghoff, U.M. (2006), Long-term preservation of digital documents: principles and practices, Springer Verlag.

Conway, E., Matthews, B., Giaretta, D., Lambert, S., Draper, N. and Wilson, M. (2011), "Managing risks in the preservation of research data with preservation networks", 7th International Data Curation Conference, Bristol, UK, December 2011.

Conway, E. and Lambert, S. (2011), "Information pack on the research datasets testbed", SCAPE: SCAlable Preservation Environments, STFC, Oxford.

CCSDS (Consultative Committee for Space Data Systems) (2002). Reference Model for anOpen Archival Information System (OAIS). Recommendation for space data system standards, CCSDS 650.0-B-1. http://public.ccsds.org/publications/archive/650x0b1s.pdf (accessed 7th August 2012).

Factor, M., Henis, E., Naor, D., Rabinovici-Cohen, S., Reshef, P., Ronen, S., Michetti, G. and Guercio, M. (2009), "Authenticity and provenance in long term digital preservation: modelling and implementation in preservation aware storage", Proceedings of the USENIX First Workshop on the Theory and Practice of Provenance (TaPP), San Francisco, USA, February.

Lee, K.H., Slattery, O., Lu, R., Tang, X. and Mccrary, V. (2002), "The state of the art and practice in digital preservation", Journal of Research-National Institute of Standards and Technology, Vol. 107, No. 1, pp. 93-106.

STFC (2012), "ISIS: A world centre for neutrons and muons", available at http://www.isis.stfc.ac.uk/ (accessed 7th August 2012).