Proceedings of the 11th International Conference on Manufacturing Research (ICMR2013), Cranfield University, UK, 19th – 20th September 2013, pp 213-218

TOWARDS A FRAMEWORK FOR PREDICTING WHOLE LIFE-CYCLE COST FOR LONG-TERM DIGITAL PRESERVATION

Mohamed Badawy
Manufacturing and Materials Department
Cranfield University
Cranfield
Bedfordshire, MK43 0AL, UK
m.badawy@cranfield.ac.uk

Essam Shehab
Manufacturing and Materials Department
Cranfield University
Cranfield
Bedfordshire, MK43 0AL, UK
e.shehab@cranfield.ac.uk

Isaac Sanya
Manufacturing and Materials Department
Cranfield University
Cranfield
Bedfordshire, MK43 0AL, UK
i.o.sanya@cranfield.ac.uk

Paul Baguley
Manufacturing and Materials Department
Cranfield University
Cranfield
Bedfordshire, MK43 0AL, UK
p.baguley@cranfield.ac.uk

ABSTRACT

Estimating the costs for the whole lifecycle of long-term digital preservation (LTDP) activities enables decision makers to choose carefully what data to preserve, duration of preservation and type of preservation techniques best applied for their information. To address this need, a framework is developed to generate a cost model that will estimate costs for long-term digital preservation activities using storage in the cloud and taking into consideration the impact of mitigating uncertainties, especially obsolescence issues on future costs. This cost estimating framework is part of the European project entitled 'Enabling kNowledge Sustainability Usability and Recovery for Economic value' which aims to provide a total long-term digital preservation solution for companies and public sector organisations interested in keeping their digital information alive for the long-term within the healthcare, financial and the clinical trials business sectors.

Keywords: Digital Preservation, Cloud Computing, Uncertainties, Obsolescence, Cost Estimation, Whole Lifecycle Cost, ENSURE.

1 INTRODUCTION

Digital data is now dominating the information scene (Charles Beagrie Limited, 2010); most of the currently generated information and records are in digital format or end up being digitized. This is due to the existence of new technologies and the ease of handling these digital data forms by the users (Hughes, 2004). In order to keep these vastly growing population of digital materials safe, useable, meaningful and accessible for a long period of time; preservation methods and techniques have been introduced and employed.

Enabling kNowledge Sustainability Usability and Recovery for Economic value (ENSURE) aims to provide a total long-term digital preservation solution for a new sector in the IT market. The businesses in healthcare, financial and clinical trials sectors are now interested to preserve their data, due to legal obligations and due to the increasing cost of data regeneration, especially in the clinical trials sector. Along with the new business sector, ENSURE is aiming to utilise cloud computing for storage and computing.

ENSURE aims to provide its customers from the three business sectors with a detailed cost and economic performance report. This will enable decision makers to select their long-term digital preservation requirements and have the highest quality preservation possible for the cheapest running

cost and ensure ease of access to and security of data. ENSURE's cost model aims to tackle uncertainties and obsolescence issues that may arise. This is due to IT systems that are prone to failures and obsolescence. Failures and obsolesce issues will generate cost to mitigate, and a rigorous study is need to estimate the impact on future cost due to these mitigation strategies. To have a real cost estimate, that reflects the effect of uncertainties, ENSURE requires that the cost modelling development should include a thorough uncertainty study. It was identified by the authors that a contribution could be considered effective if a framework is generated, that when utilised can generate a cost model suitable for long-term digital preservation. This will benefit different business sectors and enable them to make better and accurate decisions about committing to any long-term digital preservation activity. This paper outlines an initial cost-estimating framework for LTDP in cloud computing and presents the results thus far.

2 COST MODELS FOR DIGITAL PRESERVATION

Many projects containing cost models to estimate costs of digital preservation exist for specific sectors, but most of them targets the libraries, national archives, representing the heritage sector and for laboratories and research facilities, representing science facilities sector. These projects have four main cost models, NASA's Cost Estimation Tool (CET) (Hendley, 1998), Lifecycle Information for E-Literature (LIFE) (Wheatley, et. al, 2009), Cost Model for Digital Preservation (CMDP) (Kejser, et. al, 2011) and Keeping Research Data Safe (KRDS) (Stanger, 2011).

CET is the oldest, targeting the scientific sector. It was developed to estimate lifecycle costs of maintaining scientific data centres. Its tool has a comparable database of historic data reachable through a set of what-if choices and parameter sensitive tests. CET has a strong cost model with comparable databases. This database started with 29 projects data, and the model is constantly updating and adding new ones to them. A highly recognised cost model that served the heritage sector is the LIFE project, which was developed by the cooperation of University of London and British library on three phases, LIFE1, LIFE2 and LIFE3. Its main target sector is to serve libraries. The LIFE cost model looks at the complexity of file formats which it divides into 10 separate complexity levels. LIFE depends hugely on the Open Archival Information System (OAIS) reference model (CCSD, 2002), which helps in breaking down the process levels.

Another cost model that was dedicated for the science facilities sector is KRDS, which was developed by the consultancy firm Charles Beagrie Ltd (Beagrie, 2008). The project finished in 2011, and its main concern is costing for research data preservation. Based on similar projects, CET and LIFE costs data collection from multiple UK universities and a number of projects and archives. The model analyses data and develops a cost benefit relation of preservation for given data sets. KRDS strong point is that it integrates the best of the LIFE and CET, where it gets the cost benefit and the lifecycle costing from. However, due to its limitation to the science facilities sector, the LIFE flexibility is lost. Another weakness in KRDS is that it fails to provide significant details in the activities based on the OAIS reference model, unlike LIFE and CET (Stanger, 2011).

The Danish National Archive developed in 2011 a cost model, focused on two main stages, CMDP1 and CMDP2. CMDP1 is for Preservation Planning and Digital Migration and CMDP2 is focused on the ingest phase (CMDP, 2010). The model is based on the OAIS reference model and uses the activity based modelling technique (Kejser, et. al, 2011).

The previous cost models have some weaknesses in common. Firstly, they do not have any uncertainties study integrated in their cost model, which questions the practicality of the model for the future. The second weakness is that they did not take in consideration cloud computing as a storage and computing option, which for the time being has high potential because of its ease of access, scalability and fast set-up time.

3 RESEARCH METHODOLOGY

To achieve the requirements and targets of the model, a research methodology was put in place to maximise the time dedicated for this project. Figure 1 shows the three main phases of the research methodology and related activities.

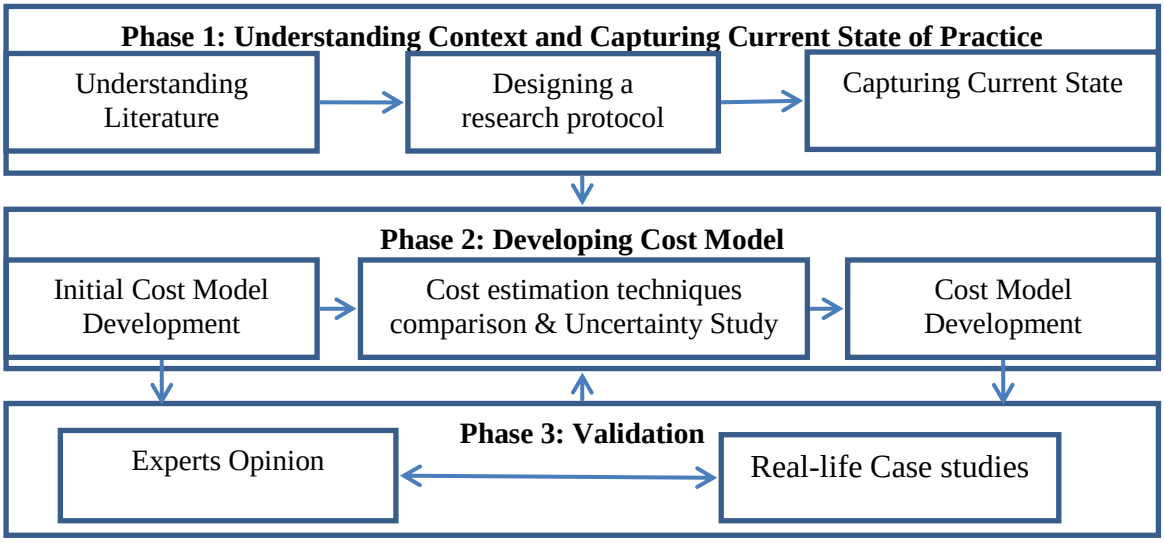

Figure 1: ENSURE Cost Model's Research Methodology (Badawy, M. et al. 2012)

Three main phases constructs the research methodology, phase one focuses on understanding state-of-the art of the science of digital preservation, cost modelling, and then capturing the current state of practice in the industry. The AS-IS model is captured using a surveying technique, in this case a questionnaire. Phase two initiates with developing an initial cost model that can represent costs for a single form of digital preservation, migration, as a starting point. Finally, phase three validates the developed cost model through experts' opinion and case studies.

4 LTDP COST PREDICTION FRAMEWORK DESIGN

The main target of this framework is to generate a cost model that can predict costs of carrying out LTDP activities in healthcare, financial and clinical trials business sectors, while considering the impact of uncertainties and obsolescence issues on estimated costs. The process of developing a framework for whole lifecycle cost of long-term digital preservation started with developing a framework to start the research with some viable information. This was followed by capturing LTDP activities for the ENSURE project. Capturing the use-case owners preservation requirements followed, a detailed Work Breakdown Structure (WBS) and Cost Breakdown Structure (CBS) of the ENSURE system and a complete Activity Based Costing (ABC) cost model was developed along with a cloud storage initial model. The cloud storage is now being expanded into a full cloud computing cost model with uncertainties and obsolescence impact on cloud computing utilisation costs, and when finished it will be implemented within the current cost model.

4.1 Framework

The framework design, as shown in Figure 2, is designed to study digital preservation whole lifecycle activities, and then generate a work breakdown structure of the current digital preservation activities. Afterwards it ensures that the activities are broken-down to their simplest forms and then choses a suitable cost estimation technique to generate a cost model. Some rules are developed for the controlled behaviour of the model, then estimation and validation of the model is carried out through experts' opinions and case studies. After the validation process, any errors or deviations is fed into the design phase of the cost model generation to amend any known issues.

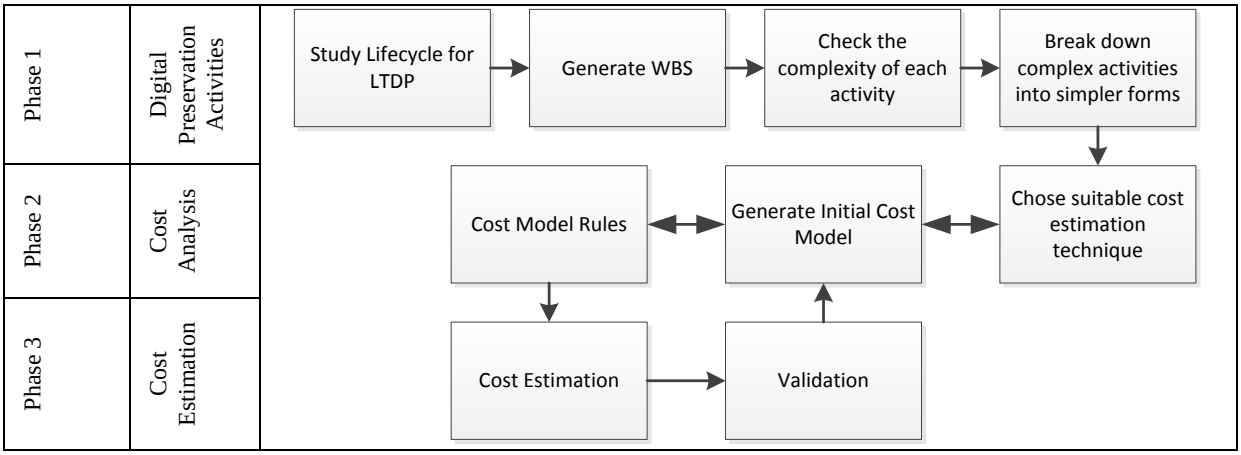

Figure 2: Initial Cost Estimating Framework for Long-Term Digital Preservation Activities

The framework was designed for the user to start with understanding the business requirements, followed by a work breakdown structure. Then studying the complexity of each work activity and make sure that it is represented in its simplest form; then from historical cost data present, the user selects a suitable cost modelling technique.

4.2 Lifecycle of Long-Term Digital Preservation

After the generation of the framework, it was essential to understand the whole lifecycle behaviour of the ENSURE preservation system, shown in Figure 3. This was achieved by carefully studying the ENSURE system requirements documents, attending the weekly system configuration meetings and carrying out two online workshops with consortium members.

Figure 3: Whole Lifecycle of LTDP

4.3 Sector Differences and Requirements

A series of interviews with use-case owners were conducted to capture their requirements for a digital preservation system. The focus was to find what is their expectation rather than their AS-IS situation. This is due to their lack of experience in the area. The main areas of differences were captured. The main areas of differences were in the preservation duration (i.e. data retention period), file types, file formats, their access rates to preserved files, copy rights issues and different legal requirements. These requirements affects the total cost of utilising the ENSURE system, for example, legal and copy rights issues will call for higher security, thus increase cost of encryption and decryption processing. Also higher access rates will generate a cost difference with relation to the storage facility used. And finally longer preservation periods will also result in an increase of incurred cost.

4.4 Work and Cost Breakdown Structure

The work breakdown structure, as illustrated in Figure 4, consists of five main areas that build together ENSURE preservation behaviour. In this WBS, Data Management is always active and continuously carrying out data checks to ensure the integrity and security of the preserved information.

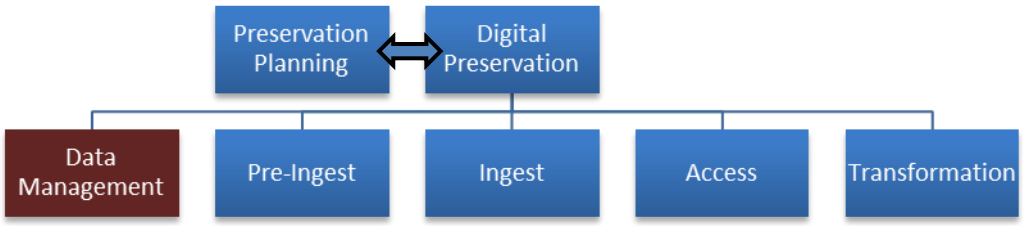

Figure 4: High-level View of LTDP WBS

The cost breakdown structure for the ENSURE system is presented in Figure 5 and shows a high level view of the cost generating activities. These activities' costs are calculated based on equations and rules. The expected outputs of the cost model are illustrated in Figure 5.

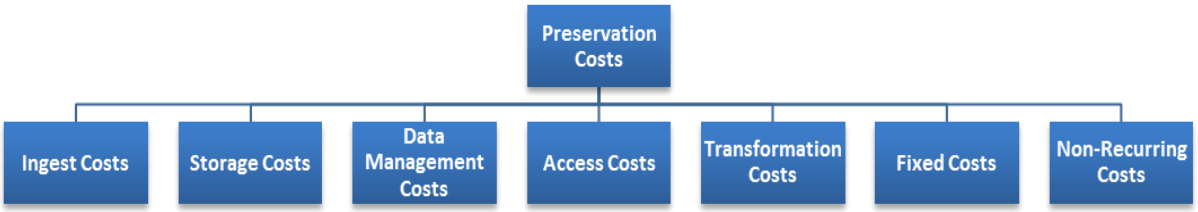

Figure 5: High level CBS for the ENSURE System

4.5 Activity Based Cost Model

The cost model has been developed into a detailed model based on the ABC estimation technique. This is part of the final expected cost model, when combined with another two separate techniques, AHP based and Parametric. Both AHP and Parametric models are under development. Mathematical equations were developed based on ENSURE's system activities that have been identified within the WBS. The cost model takes public cloud storage facilities and private OpenStack based storage options. It is expected that the total preservation cost is going to be represented by the cost of ingesting, the cost of archiving and the cost of accessing the data. The following formulae are the high-level formulae of the developed cost model.

- $Ingest\ Cost = Information\ Package\ generation\ cost + Quality\ check\ cost + metadata\ generation\ cost + Data\ protection\ for\ ingest\ cost$
- $Storage\ Cost = Data\ Transfer\ Cost + Storage\ cost$
- $Data\ Management\ Cost = Fixity\ check\ cost + Reporting\ cost + File\ deletion\ cost + amendment\ to\ metadata\ cost + access\ audits\ cost$
- $Access\ Cost = Information\ Package\ retrieval\ cost + Data\ Protection\ for\ Access\ cost$
- $Transformation\ Cost = Data\ Migration\ cost + Virtual\ Appliance\ initiation\ cost$

5 INITIAL VALIDATION

A dual validation is designed for this research project, Qualitative and Quantitative. For the initial stage of the research most of the validation is done qualitatively through experts' opinion. Weekly consortium conference call meetings and regular online workshops are carried out, to validate both breakdown structures, WBS and CBS, the whole lifecycle of LTDP and sector differences. When finalising the framework, quantitative validation methods will be employed to validate the framework's output and behaviour. This will be done through at least three case studies.

6 CONCLUSIONS

Cost models for LTDP that have been developed over time has primarily focused to serve the library or the scientific centres communities; so all activity studies, case studies and breakdown structures are most suitable to these business sectors This makes it clear that no work has been done for developing a cost model for LTDP for the financial sector, healthcare sector and clinical trials sector. These sectors are highly interested in digital preservation for two main reasons. Primarily that it is a legal obligation on all of them to keep their data for at least 5 years, some up to 30 years and even more. The other important reason is that these businesses do not want to lose valuable data they paid to generate since it will cost them higher to acquire again. None of the existing cost models attempted to calculate costs for cloud computing utilisation in digital preservation; while it is known that cloud computing should decrease the initial costs and the effect of obsolescence on cost. Another critical point in this discussion is risk, uncertainty and obsolescence in digital preservation. Previous cost models did not take into account the cost of mitigating obsolescence. This is a major drawback for previous long-term digital preservation cost models, because the idea of being long-term effective is mitigating obsolescence. Thus it is highly important to investigate what are the actual issues of obsolescence facing long-term digital preservation. Then investigate the possibility of including the cost impact of mitigation activities within the cost model being researched. Finally the research results provide a solid step towards the final development of the cost estimating framework for LTDP activities.

ACKNOWLEDGMENTS

The research has been conducted as part of an EU-FP7 project titled ‘ENSURE'. The project involves 13 research and industrial partners. The 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

Badawy, M., Shehab, E., Baguley, P. and Wilson, E. (2012), “Towards a Cost Model for Long-Term Digital Preservation”, ISPA/SCEA Joint International Conference & Training Workshop, 14th – 16th of May 2012, Brussels

Beagrie, N., Chruszcz, J. and Lavoie, B. (2008), Keeping Research Data Safe: A Cost Model and Guidance for UK Universities, , Charles Beagrie Limited, UK.

CCSDS (2002), Reference Model for an Open Archival Information System (OAIS), CCSDS 650.0-B-1, Consultative Committee for Space Data Systems, http://public.ccsds.org/publications/archive/650x0b1.PDF (last visited 25/06/2011).

CMDP Official Website, 2010, http://www.costmodelfordigitalpreservation.dk/ (Last visited 26/01/2012)

Hendley, T. (1998). Comparison of methods and costs of digital preservation. A JISC/NPO Study within the Electronic Libraries (eLib) Programme on the Preservation of Electronic Materials

Kejser, U. B., Nielsen, A. B., & Thirifays, A. (2011). Cost model for digital preservation: Cost of digital migration. International Journal of Digital Curation, 6(1), p 255-267

Russell, K., & Weinberger, E. (2000). Cost elements of digital preservation (Online Draft), http://www.scribd.com/doc/7345161/RUSSELL-Kelly-Cost-elements-of-digital-preservation, (last visited 09/01/2012)

Stanger, N. (2011). Keeping research data safe. Computer and Information Science Seminar Series, http://otago.ourarchive.ac.nz/handle/10523/1502, [Accessed on 23 May 2011].

Wheatley, P. and Hole, B. (2009). LIFE3: Predicting long-term digital preservation costs. http://www.life.ac.uk/3/docs/ipres2009v24.pdf (last visited 17/01/2012)