Proceedings of the 11th International Conference on Manufacturing Research (ICMR2013), Cranfield University, UK, 19th – 20th September 2013, pp 183-188

A COST ENGINE SYSTEM FOR ESTIMATING WHOLE-LIFE CYCLE COST OF LONG-TERM DIGITAL PRESERVATION ACTIVITIES

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

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

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

ABSTRACT

This research paper presents a cost engine system that estimates the whole life cycle cost of long-term digital preservation (LTDP) activities using cloud-based technologies. A qualitative research methodology has been employed and the activity based costing (ABC) technique has been used to develop the cost model. The unified modelling language (UML) notation and the object oriented paradigm (OOP) are utilised to design the architecture of the software system. In addition, the service oriented architecture (SOA) style has been used to deploy the function of the cost engine as a web service in order to ensure its accessibility over the web. The cost engine is a module that is part of a larger digital preservation system and has been validated qualitatively through experts' opinion. Its benefits are realised in the accurate and detailed estimation of cost for companies wishing to employ LTDP activities.

Keywords: cost engine, long-term digital preservation, whole life cycle cost, cloud computing.

1 INTRODUCTION

Digital preservation is defined as a series of activities that enables the continuous access to digital information for as long as required. It involves the planning and application of preservation techniques and technologies to ensure the active management and accessibility of digital information. Becker et al (2009) suggests that digital preservation is more complex than traditional preservation. Unlike traditional objects (e.g. photographs, books) where users can immediately access its content, digital objects are dependent on both the hardware and software environment. However, these environments are subjected to rapid evolvement due to technological obsolescence which threatens continuous access to digital content. Estimating the accurate cost for preserving digital information has become a challenge due to the need to account for uncertainties and obsolescence issues on cost. However, understanding the cost for LTDP activities will enable companies to make better decisions in selecting the most appropriate preservation solution from a cost perspective.

In recent times, cloud computing is rapidly becoming an interesting technique for supporting digital preservation. With easy access from any location and significant reduction of initial investment costs, cloud computing is gradually becoming a preferred preservation approach (Badawy et al. 2012). Cloud computing is described as the utilisation of computing resources that are distributed over a network (i.e. the Internet) or a service. End users accesses cloud-based applications via a light-weight

application and user's data are stored on remote servers. There are many types of cloud deployment models such as public cloud, private cloud, hybrid cloud and community cloud. All has their advantages and disadvantages and companies should carefully study the cost/economic implications of these cloud deployment models before making decisions on LTDP projects. It is important to identify that cloud computing is considered as an enabler of digital preservation (i.e. storage option) and it is not on its own enough for performing LTDP activities.

Enabling k**N**owledge **S**ustainability, **U**sability and **R**ecovery for **E**conomic Value (ENSURE) is an EU-FP7 (Framework Programme Seven) project focused on providing a total solution for LTDP. The ENSURE system is focused on serving three main business sectors. These are namely financial, healthcare and clinical trials. Previous LTDP initiatives has traditionally been geared towards the scientific and heritage sectors. Thus, the aforementioned sectors ENSURE is targeting is unique and so is the cost engine system being developed for this purpose. The ENSURE system is comprised of three main engines, the first is the cost engine (CE), followed by the economic performance assessment engine (EPAE) and lastly, the quality engine (QE). These engines are integrated into the larger architecture of the ENSURE system as illustrated in Figure 1. The purpose of these engines are to provide crucial decision-support to users wishing to employ LTDP activities. This includes evaluating and optimising the cost and benefit of many preservation plans to determine the most appropriate in terms of cost and quality. The ENSURE architecture has several modules communicating together with a wide variety of functionalities.

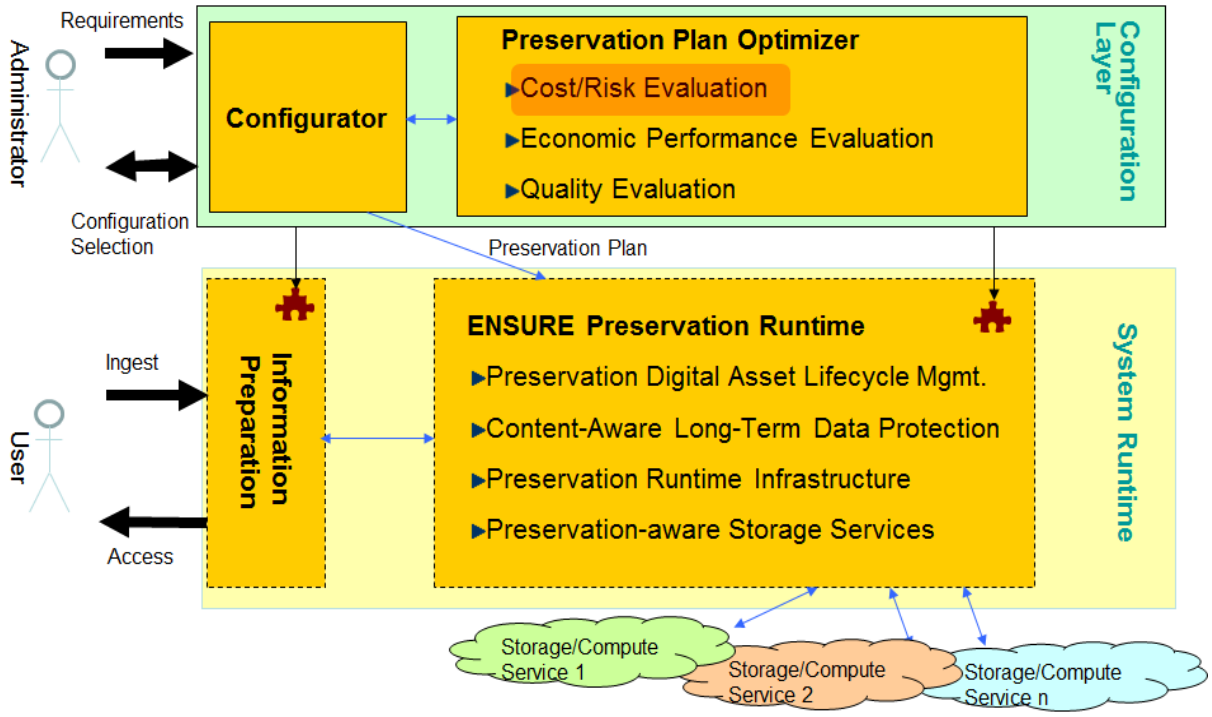

Figure 1: ENSURE System Architecture (ENSURE 2013)

As illustrated in Figure 1, the cost engine is a module within the configurator layer of the ENSURE system. This research paper describes the cost engine module of the ENSURE system and outlines how it can be used as a crucial decision-support mechanism for estimating costs for LTDP activities on the cloud.

Although there have been many attempts to develop cost models within LTDP, these cost estimating systems are usually not described in detail and often do not cover the 'whole life cycle cost' of LTDP. Some are focused on storage only, others are focused on other aspects of LTDP activities (e.g. ingest, data management). Another limitation is that most of these cost systems are geared towards the scientific and heritage sectors. There is lack of research focused on developing cost estimation systems for LTDP activities for the financial, healthcare and clinical trials sector. Thus, this research paper presents a cost engine system that aims to estimate the whole life cycle cost

for LTDP activities for the financial, healthcare and clinical trials sector in order to enable crucial decision-support for users in these sectors.

2 RELATED WORK

There are many projects focused on developing cost models for estimating LTDP costs. However, most of these cost models are geared towards the libraries, national archives and science facilities sectors. There are four main LTDP cost models which are namely 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 lastly, Keeping Research Data Safe (KRDS) (Stanger 2011).

NASA CET is focused on targeting the scientific sector. It was designed to estimate the lifecycle costs involved in maintaining and managing data centres. The tool is based on historical data and based on 'what-if' scenarios. Another cost model for LTDP is the LIFE project which consists of three phases, namely LIFE1, LIFE2 and LIFE3. The main target and focus of the LIFE project is to serve the library and heritage sectors. The LIFE cost model focuses on examining file complexity and it depends mainly on the open archival information system (OAIS) standard model (CCSDS 2012) which is used to breakdown the cost.

KRDS LTDP cost model was developed to serve the science facilities sector. The model is similar to NASA CET and LIFE in that its data comes from multiple UK universities and a number of archives and projects. The cost model is used to assess a cost benefit analysis of preserved data. KRDS mainly integrates the best functions of the NASA CET and LIFE project. However, its limitations with the science facilities sector questions its applicability to other sectors. A limitation of KRDS is also that it fails to define the cost breakdown and its relation to the OAIS standard model (Stanger 2011).

CMDP was developed by the Danish National Archive and has two main phases. CMDP1 is focused on preservation planning and information migration while CMDP2 is focused solely on the ingest phase (CMDP 2010) and not the whole lifecycle cost. The cost model uses the activity based costing (ABC) approach and is closely related to the OAIS model.

The literature has highlighted that there have been many attempts to accurately estimate cost for LTDP activities. However, these attempts are mainly focused towards the scientific and heritage sectors. There is lack of research investigating the development of LTDP cost estimation systems for the financial, clinical trials and healthcare sectors. Thus, this research study presents the design and development of a cost engine system that is fully integrated into a large digital preservation system and aims to estimate the whole life cycle cost for preserving digital information using cloud computing for the aforementioned sectors.

3 ADOPTED RESEARCH METHODOLOGY

Due to the collaborative nature of the research study, a research methodology satisfying both industry and academia has been selected. Both qualitative and quantitative research approach has been used to design and develop the ENSURE cost engine system. An activity-based costing approach has been employed to design the software system. In the first phase of the research method, the identification of cost drivers and cost break down structures (CBS) was completed and validated with seven digital preservation experts using a series of interviews and discussions. Following on from this, over 40 equations and rules were derived representing costs of LTDP activities. The next phase was to design the system using the UML class diagram to represent cost objects in order to elicit cost metrics for each object. Attributes of each cost objects were represented followed by identifying the relationships between the cost objects. An architecture diagram was designed to represent how the cost engine interacts with the rest of the ENSURE system. This activity involved clearly identifying inputs from the ENSURE configurator and outputs of the cost engine system for LTDP activities. Furthermore, the service oriented architecture (SOA) approach was used to design the cost engine system as a web service in order to deploy its functions publicly and as an open-source technology. Lastly, the cost engine system business logic was implemented using the object oriented (OO) java programming language containing over 2000 lines of code-logic. The cost engine is now fully integrated as part of the ENSURE system and has been validated qualitatively through experts opinion. Quantitative

validation is underway and the cost engine system will be validated with experts and use-case partners within the financial, healthcare and clinical trials sector.

4 COST ENGINE EQUATIONS AND UML DESIGN FOR LTDP ACTIVITIES

Table 1 illustrates an example of some of the cost engine implemented equations. The cost represented in this equations are namely, data management fixity check cost, cloud storage cost, ingest cost and data protection encryption cost.

Table 1: Excerpts of Cost Engine Implemented Calculations

Cost Metric (€)	Cost Calculations	Description
Data Management Fixity Check Cost	$\left(\left(storage\ volume\ (GB) \div fixity\ check\ processing\ rate\ \left(\frac{GB}{hr}\right)\right) * processing\ cost\ \left(\frac{€}{hr}\right)\right) * (frequency\ of\ checks\ \left(\#\frac{no}{year}\right) * data\ retention\ period)$	All costs incurred to carry out all fixity checks for a set of data for the retention period
Cloud storage cost	$\left(storage\ volume\ (GB) * volume\ price\ \left(\frac{€}{GB}\right) * storage\ duration\ (month)\right) + ((cost\ of\ requests\ (€) \times number\ (\#) of\ requests)$	All costs incurred to store the files on the cloud storage system
Information Package Generation Cost	$\left(storage\ volume\ (GB) \div ingest\ processing\ rate\ \left(\frac{GB}{hr}\right)\right) \times ingest\ processing\ cost\ \left(\frac{€}{hr}\right)$	The cost incurred to generate the SIP (Submission Information Package)
Data Protection Encryption Cost	$\left(storage\ volume\ (GB) \div encryption\ processing\ rate\ \left(\frac{GB}{hr}\right)\right) * encryption\ processing\ cost\ \left(\frac{€}{hr}\right)$	All costs incurred to carry out data protection actions on ingested files

Figure 2 illustrates the OOP UML cost engine design. The cost engine classes are defined by the core activities carried out in digital preservation. These are ingest, access, storage and data management activities. Transformation and staff costs are also included to cover a wider range of cost drivers. Furthermore, it has been identified that companies wishing to carry out digital preservation activities would like to know how much they need to invest initially to start up preservation activities. Thus, an initial investment cost object was developed in the cost engine model to describe non-recurring cost (e.g. virus software cost) and infrastructure cost (e.g. physical space cost, cooling cost for servers).

5 ENSURE COST ENGINE SYSTEM ARCHITECTURE

The cost engine system architecture is comprised of several communicating components that implement the overall ENSURE cost evaluation and optimisation system. Figure 3 illustrates the architecture of the cost engine and how its modules interact with the rest of the ENSURE system. The Global Preservation Plan (GPP) is passed from the configurator to the preservation plan optimiser (PPO) with a full set of parameters. The GPP is comprised of the preservation plan and preservation configuration. The preservation plan describes 3 main types of actions associated with digital preservation activities. These are: system actions which affect all aggregations within the preservation plan; aggregation actions which affect specific aggregations; and copy actions which affect only one copy of a specific aggregation (e.g. Rackspace EU copy for market data aggregation). Copy actions can often be described as primary and secondary storage actions. The preservation configuration describes the physical architecture, software and plugins employed for digital preservation activities. These plugin service instances are implicitly interconnected to the actions within a preservation plan via their plugin service instances id description.

The PPO is the master component which controls the cost engine component and manages its dialog with the Configurator. Within the cost engine architecture, the cost engine web service module describes and implements the cost engine business logic for both the evaluation and optimisation operations. The PPO provides the cost engine's input parameters via the configurator. The inputs

include the GPPs to be evaluated, a list of plugin service instances (PSI), requirement sets and user preferences.

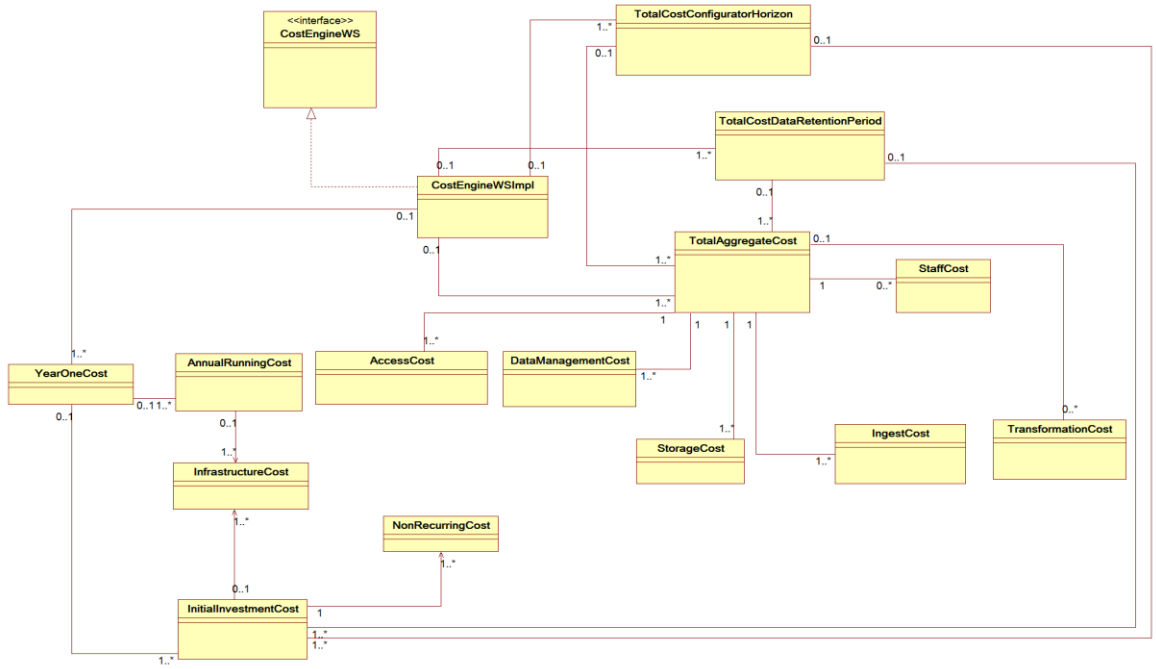

Figure 2: ENSURE UML Cost Engine Simplified Class Diagram

The data handler component manages the data flow between the configurator, the PPO and optimisation engines (i.e. cost engine, quality engine and economic performance engine). The output of the cost engine is converted into a set of java classes. These classes are pre-defined in the configurator and passed back to the PPO which then communicates the results to the end user through the ENSURE graphical user interface. The inputs and outputs of the configurator classes are explicitly integrated with the cost engine classes. Therefore, the cost engine model is translated to the configurator model to ensure common semantics between the overall ENSURE system. The Configurator Utilities library was used by the cost engine for testing and verifying the cost engine's business logic implementation. Furthermore, error handling was incorporated into the cost engine business logic in view of ensuring the security and integrity of the cost engine system.

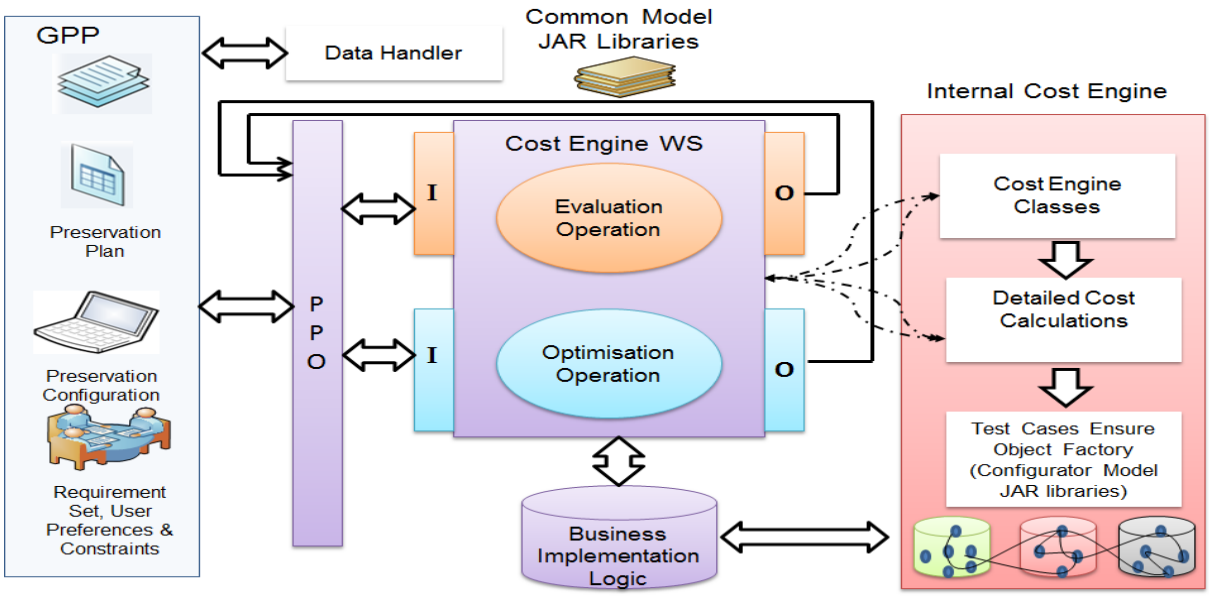

Figure 3: ENSURE Cost Engine Architecture Model Description

6 VALIDATION

The cost engine system has been validated qualitatively through experts' opinion. This has been achieved through various mediums such as discussions and weekly group meetings with the rest of the ENSURE consortium. The first phase of validation involved validating the equations and rules of the cost engine with the ENSURE consortium which involves 12 industrial partners from an IT, healthcare, clinical trials and financial industry. The second phase of validation involves the validation of the cost engine integration with the ENSURE configurator layer. Several test cases has been conducted which matches the expected outputs of the cost engine to its actual outputs. This include verification of the cost engine business logic and the cost engine interfaces. Quantitative validation will commence which will focus on validating the cost metric numbers with real numeric numbers from the use-case companies (i.e. clinical trials, healthcare and financial). The quantitative validation will consist of several case studies with use-case partners and other digital preservation experts in order to ensure the applicability, validity and generalisability of the cost engine system.

7 CONCLUSIONS

This research paper has presented the design and development of a cost engine system for estimating the whole life cycle cost of LTDP activities for the finance, healthcare and clinical trials sector. The main cost drivers for LTDP activities has been identified and translated into a set of mathematical formulas to enable the accurate estimation of LTDP cost. The cost engine has been implemented using the object oriented approach and the java programming language. Additionally, the service oriented architecture approach has been employed to develop the cost engine as a web service. The cost engine system is now fully integrated into the overall ENSURE system. The ENSURE configurator input is passed by the preservation plan optimizer to the cost engine input parameters. As a result of this, the cost engine calculations are executed which in turn outputs an estimated cost for LTDP activities for each candidate GPP. The benefit of the cost engine system is an accurate cost estimation for LTDP activities which companies belonging not only in the healthcare, clinical trials and financial sector but also manufacturing industries can use as a crucial decision support tool.

ACKNOWLEDGEMENTS

The research has been conducted as part of a 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 (http://http://ensure-fp7-plone.fe.up.pt/site).

REFERENCES

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

Becker, C, Kulovits, H, Guttenbrunner, M, Strodl, S, Rauber, A and Hofman, H. (2009). Systematic planning for digital preservation. International Journal on Digital Libraries (10): pp.133–157.

CCSDS (2012). Reference Model for an Open Archival Information System (OAIS). Available from: http://public.ccsds.org/publications/archive/650x0m2.pdf [Accessed on 17 Apr 2013]

CMDP, (2010), OAIS, http://www.costmodelfordigitalpreservation.dk/ [Accessed on 26 Jan 2012]

ENSURE. (2013). A2 Configurator Layer Work Description. Year 2 Scientific Report. http://ensure-fp7-plone.fe.up.pt/site/deliverables [Accessed on 17 Apr 2013]

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

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 [Accessed on 17 Jan 2012]