CRANFIELD UNIVERSITY

MOHAMED ABDEL SALAM ABDEL AZIM BADAWY

A FRAMEWORK FOR WHOLE LIFECYCLE COST OF LONG–TERM DIGITAL PRESERVATION

SCHOOL OF AEROSPACE, TRANSPORT AND MANUFACTURING

PhD
Academic Year: 2016 - 2017

Supervisors: Professor Essam Shehab & Dr Paul Baguley
March 2017

CRANFIELD UNIVERSITY

SCHOOL OF AEROSPACE, TRANSPORT AND MANUFACTURING

PhD

Academic Year 2016 - 2017

MOHAMED ABDEL SALAM ABDEL AZIM BADAWY

A FRAMEWORK FOR WHOLE LIFECYCLE COST OF LONG–TERM DIGITAL PRESERVATION

Supervisors: Professor Essam Shehab & Dr Paul Baguley

March 2017

This thesis is submitted in partial fulfilment of the requirements for the degree of PhD

ABSTRACT

Digital preservation, also known as digital curation, is the active management of digital information, over time, to ensure its accessibility and usability. Digital preservation is nowadays an active area of research, for many reasons: the rapid evolution of technology, which also results in the rapid obsolescence of old technologies; degradation of physical records; constantly increasing volumes of digital information and, importantly, the fact that it has started to become a legal obligation in many countries.

This research project aims to develop an innovative framework estimate costs of long term digital preservation. The framework can lead to generating a cost model that quantifies costs within different business sectors, while capturing the impact of obsolescence and uncertainties on predicted cost. Case studies from financial, healthcare and clinical trials sectors are used to prove the framework concept. Those sectors were chosen because between them they share all file types that are required to be preserved and all are either obliged by European or local laws, e.g. EU Data Retention Directive (2006/24/EC) and/or UK Data Retention Regulations 2014 No. 2042, or interested in preserving their digital assets.

The framework comprises of three phases: assessing digital preservation activities, cost analysis and expansion and cost estimation. The framework has integrated two processes that will enable the user to reach a more accurate cost estimate; a process for identifying uncertainties with digital preservation activities and a cost modelling process. In the framework cloud computing was used as an example for storage and compute technologies.

Combining different research methodology techniques was used in this research project. Starting with conducting a thorough literature review covering digital preservation and cost modelling. Following the literature review; is a combination qualitative and quantitative approaches, using semi-structured interview technique to collect data from industry experts. Industry experts were chosen from companies, firms and government bodies working with or researching digital

preservation. Finalising with validating results by real-life case studies from businesses in selected sectors and experts' verdict.

Comparing the output of the framework to real-life case studies, demonstrated how companies/firms, who target to preserve their digital assets, can utilise it to predict accurately future costs for undertaking such investment. By applying industrially-based cost modelling approaches the framework generates a cost model that predicts single-point and three-points cost estimates, an obsolescence taxonomy, uncertainties identification process and quantifying uncertainties and obsolescence impact on cost prediction. Providing decision makers with all the framework outputs, will provide them with quantifiable information about their future investment, while remaining clear to understand and easy to amend. This makes the framework provide long-term total cost prediction solution for digital preservation to firms; helping, guiding and adding insight into digital preservation added value.

Keywords: Cost Estimation, Cost Modelling, Digital Curation, Uncertainty Cost, Obsolescence Cost, Obsolescence Taxonomy

ACKNOWLEDGEMENTS

I would like to express my deep appreciation to many people who contributed to this report, either directly or indirectly, for working together to put me on the right track in this PhD research project.

Starting with my parents, back home, without whom I could not have been here to reach this level of education. Many thanks also to my wife for her patience, continuous encouragement and support. In addition, I would like to thank my sister here in the UK, for her continuous help and direct support.

I also would like to take this opportunity to show my deepest gratitude and appreciation to Professor Essam Shehab and Dr Paul Baguley, who supported me with their knowledge, patience, constructive criticism and accurate direction. I also want to deeply thank Dr Isaac Sanya, who invested his time in me and gave my project such a successful start.

I would like also to thank deeply my subject advisor, Professor Rajkumar Roy and the chairman of my review, Dr Keith Weatherhead, for spending their time and efforts to review my progress.

It has been a pleasure working in the Cranfield University's family. To all the staff who contribute to this very welcoming environment, a very warm thank you.

Finally, yet importantly, I would like also to express my appreciation to the project funding bodies, all the ENSURE project consortium and especially to Cranfield University.

All deserve from me the highest appreciation and respect.

To all who are mentioned, my deepest gratitude for their help and apologies for the inconvenience I must have inflicted.

Mohamed Badawy

LIST OF PUBLICATIONS

Peer Reviewed Conference Papers:

1. Badawy, M., Shehab, E. and Baguley, P. (2013), “Towards a Framework for Whole-Lifecycle Cost Estimation of Long-Term Digital Preservation”, 11th International Conference on Manufacturing Research ICMR 2013, 19th – 20th September 2013, Cranfield
2. Shehab E M, Chuku G and Badawy M. (2013) “A Framework for Identifying Uncertainties in Long-Term Digital Preservation” Proceedings of The 11th International Conference on Manufacturing Research (ICMR 2013), Cranfield University, Cranfield UK, 19th – 20th September 2013, pp 151-156, ISBN 978-1-907413-23-0, ISSN 2053-3373.
3. Sanya I, Shehab E, Badawy M (2013). “A Cost Engine System for Estimating Whole-Life Cycle Cost of Long-Term Digital Preservation Activities.” 11th International Conference on Manufacturing Research ICMR 2013, Cranfield, UK 19th – 20th September 2013, pp 183-188, ISBN 978-1-907413-23-0, ISSN 2053-3373.
4. Shehab E, Lefort A, Badawy M, Baguley P, Turner C, Wilson M, Conway E (2013). “Modelling Long Term Digital Preservation Costs: A Scientific Data Case Study.” 11th International Conference on Manufacturing Research ICMR 2013, Cranfield, UK 19th – 20th September 2013, pp 133-138, ISBN 978-1-907413-23-0, ISSN 2053-3373.
5. Shehab E, Sanya I, Badawy M, Ocal F, Morineau J, Fernandez Ortiz Z, Odika S and Namiesnik B (2013). “Cost Modelling for Cloud Computing Utilisation in Long Term Digital Preservation.” 11th International Conference on Manufacturing Research ICMR 2013, Cranfield, UK 19th – 20th September 2013, pp 450-455, ISBN 978-1-907413-23-0, ISSN 2053-3373.
6. 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

7. Xue, P., Badawy, M., Shehab, E. and Baguley, P. (2011), "Cost Modelling for Long–Term Digital Preservation: Challenges and Issues", 9th International Conference on Manufacturing Research ICMR 2011, 6th – 8th of September 2011, Glasgow
8. Shehab, E., Thomassin, M. and Badawy, M. (2011), "Towards a Cost Modelling Framework for Outsourcing ERP Systems", 18th ISPE International Conference on Concurrent Engineering, Advanced Concurrent Engineering, MIT, 9th August 2011, Part 6, pp. 401-408

Articles:

9. Badawy, M., Shehab, E. (2013), “ACostE Annual Conference 2012 – Cost Matter” Article, Coventry, The Journal of the Association of Cost Engineers, Vol. 51, February2013.
10. Badawy, M., Shehab, E., Turner, C. (2012) “How future safe are your digital assets and how can you ENSURE their safety?”, The Journal of the Association of Cost Engineers, Vol. 50, No. 5, September 2012.
11. Badawy, M. (2012) “2012 ISPA/SCEA Joint International Conference & Training Workshop Article”, The Journal of the Association of Cost Engineers, Vol. 50, No. 5, September 2012.
12. Erkoyuncu, J., Badawy, M. (2011), “ACostE Annual Conference 2011 Article”, ACostE Annual Conference 2011, Birmingham.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CDB	Comparables Database
CET	Cost Estimation Toolkit
DP	Digital Preservation
DRAMBORA	Digital Repository Audit Method Based On Risk Assessment
ENSURE	Enabling kNowledge Sustainability Usability and Recovery for Economic value
IT	Information Technology
KRDS	Keeping Research Data Safe
LTDP	Long-Term Digital Preservation
LIFE	Lifecycle Information for E-literature
OAIS	Open Archival Information System
NASA	National Aeronautics and Space Administration
NESTOR	Network of Expertise in long-term STOrage of digital Resources
PREMIS	Preservation Metadata: Implementation Strategies
R&D	Research and Development
TRAC	Trustworthy Repositories Audit & Certification: Criteria and Checklist

1 INTRODUCTION

1.1 Background

Natural degradation of records in different formats, the rapid increase in the volume of digital data generated, the constant evolution of technology and today's legal obligations may be the main drivers of the comprehensive research now being done around digital preservation. The rapid evolution of digital preservation technologies and systems, and the nature of managing these systems generate the need to estimate the cost of long-term preservation. The need to preserve information contained in digital assets arises from fear of losing access to information contained within each file or digital object. The fragile nature of digital information is also contributing to the problem because a minor change in a set of binaries can and will affect a file's integrity. Loss of access to information can also be caused by the dependence of digital information on specific technologies; whether these technologies are hardware or software based.

This rapid evolution of different technologies makes previous technologies obsolete, when certain conditions prevail. For example, if a new technology creates no significant change or benefit beyond the existing one, then the latter is not outdated. In contrast, if an emerging technology provides technical and cost improvement, this becomes the perfect scenario for existing technology to become obsolete. Consequently, digital preservation techniques are needed as mitigation strategies. Ensuring future access to digital assets, justifies the cost of working on digital preservation.

For a decision maker, it is crucial to understand the firm's financial commitments for the foreseeable future, and since long-term preservation investments extend over many years, the need for clearer cost understanding if anything increases. This understanding is relevant to any obsolescence issues that may arise and to other uncertainties that can impact on performance or finance availability for Long Term Digital Preservation (LTDP) systems. Estimating the future cost behaviour for an LTDP system will enable decision makers to pin-point which digital assets to preserve and which to ignore or have a less strict access policy for. Since a clear financial plan can be drawn, presenting the case for investing in LTDP becomes more viable and might help choosing compute and storage technologies.

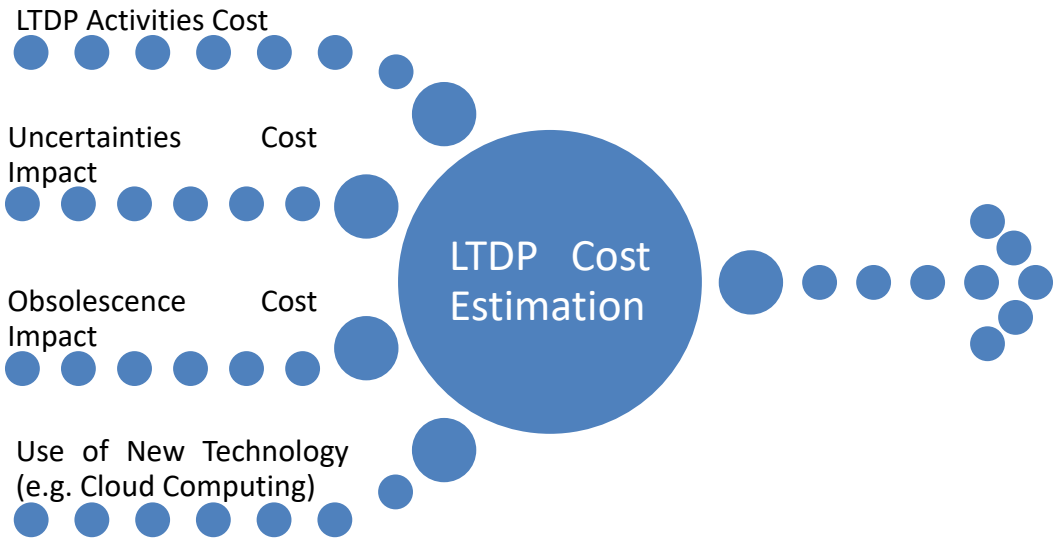


Figure 1-1 LTDP Cost Output Composition

Generating an accurate cost estimate of a digital preservation system (**Figure 1-1**) can be challenging for the system user. These challenges arise from the concept of digital preservation itself, where obsolescence, the main reason for

preservation (Kuny, 1998), is unexpected and its probability and impact is very uncertain. Regulations of a country, inter-company rules and technological advancements might engender the difficulty in predicting costs for LTDP systems (Xue, et. al, 2011). Uncertainty increases especially when state-of the-art technologies are utilised in LTDP systems, e.g. cloud computing; since these technologies obsolescence isn't presently foreseen.

Uncertainties are unknown risks and/or opportunities whose frequency and impact are not fully known; thus Obsolescence, as an uncertainty, could impact uncontrollably (Romero Rojo, 2011). This generates several cost prediction difficulties. Along with applying LTDP to new business sectors and using cloud computing, the cost estimation task should be innovative.

1.2 Problem Statement and Research Challenges

The main problem is to identify how industrially-based cost modelling approaches will be employed in different areas of digital preservation, especially long–term preservation activities; it is hard to predict cost, given the impact of uncertainties and obsolescence on these activities. After this the existing cost modelling methods and techniques needs to be examined if it's possible to represent the entire lifecycle of digital preservation. Challenges arise from the fact that a detailed cost model for businesses has never been attempted before and could encounter multiple security hazards and uncertainties.

Clear targets must be established in approaching a solution for the research project, to pave the way to a clearer definition of the problem. After establishing a cost modelling technique, the study considered ways of identifying the cost

drivers for long–term digital preservation and cloud computing use for this scenario, that will make it possible to construct a reliable cost model.

The study addresses what the uncertainties and especially the obsolescence issues are in the lifecycle of digital preservation. Obsolescence issues are significant in long-term digital preservation research, because they are the reason that preservation is needed. Using cloud computing will impact on the cost of uncertainties and obsolescence, since cloud computing is different from the traditional computing methods used to preserve data.

The last question is whether the framework generated for cost prediction going to be sector-specific or whether it could be expanded to a generic framework for digital preservation. If it can be expanded, its use in other business sectors with a few changes or tweaks will be made much easier.

The challenges are found in four main areas: technological, methodological, related to the availability of cost information and business-related.

- Technological

Technological challenges come from rapid developments in technology and the amount of detail and variety in current technologies. The use of cloud computing presents a new path for research, along with questions and difficulties.

- Methodological

The methodology of research is challenged by ensuring a validation process that is not biased and can cover the research output. The development of research results must cover all the required tasks and handle all the areas of innovation.

- Availability of Cost Information

Most of the research area is new; this threatens the availability of cost information. It is only increased if firms and companies in the sectors under investigation require some information to be withheld.

- Business sector bound challenges

Business sector differences may generate challenges to research outcomes, along with individual firms'/companies' regulations, which may impact an entire sector.

1.3 Research Questions

- Can a conceptual framework be developed to enable LTDP systems users to develop a cost model that predicts the cost of using such systems?
- How to adapt existing cost modelling methods and techniques to the entire lifecycle of digital preservation?
- How to identify the cost drivers for long–term digital preservation and what are they?
- What are the uncertainties incorporated in life cycle of digital preservation and can their impact on cost prediction be quantified?
- What are the obsolescence issues incorporated in life cycle of digital preservation and can their impact on cost prediction be quantified?

1.4 Parent Project "ENSURE"

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 ever-growing preservation market. The businesses in healthcare, finance and clinical trials are now interested in preserving their data, due to legal obligations and to the increasing cost of data regeneration, especially in the clinical trials sector. Along with new business sectors, ENSURE is aiming

to lead cloud computing in its storage methods and use the computing power of the cloud providers.

The interest of financial, healthcare and clinical trials firms in long-term digital preservation drives the work of the “ENSURE” project to find a means of better understanding the cost prediction of digital preservation for them. This is because until early 2011 there were few preservation projects that develops an LTDP cost model for business users or targeted at harnessing the capacity of cloud computing to store and actively manage digital content. This was one of the key drivers for the “ENSURE” cost modelling framework.

Therefore, to satisfy these emerging needs, this research project will employ industrially-based approaches of cost modelling to tackle issues for long-term digital preservation. Along with providing enough information about uncertainties and obsolescence, mitigations and implications, this information will help to construct a supporting tool for solving preservation issues utilising the cloud.

ENSURE aims to provide its customers from the three business sectors mentioned above with a full report on cost and economic performance. This will enable decision makers to optimise their long-term digital preservation needs, secure the highest possible quality of preservation for the cheapest running cost and ensure ease of access to their data, kept as safe as they require.

ENSURE’s cost model aims to be ready for any uncertainties and obsolescence issues. It is necessary because IT systems are prone to failures and obsolescence. These issues always generate cost through mitigating the expected effects, if these effects are estimated rigorously. To have a robust cost

estimate, which takes account of the effect of uncertainties, ENSURE requires the cost modelling development to include a thorough uncertainty study.

Thirteen consortium members are the force behind ENSURE, all of them contributing to attaining its targets.

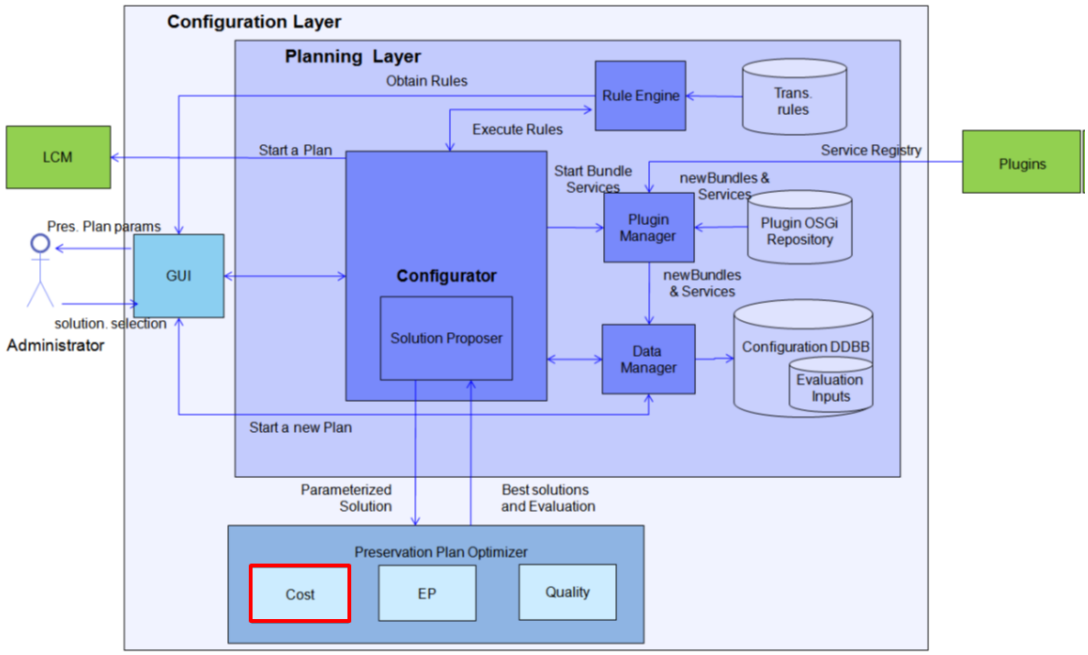


Figure 1-2 ENSURE Configuration Layer with the Cost Engine Highlighted (ENSURE, 2012)

The cost model requirements are to be able to estimate long-term digital preservation costs while keeping in mind that the system will use cloud computing. The estimated cost should integrate uncertainties and especially obsolescence factors in the cost estimation, while providing a process for prioritizing uncertainties.

Figure **1-2** shows the designed location of the cost model within the configuration layer of ENSURE's system. It shows the cost model in the Preservation Plan Optimiser, and will interact with the Quality and Economic Performance (EP)

engines. This optimiser will be "*in charge of selecting the best preservation solutions given the evaluations provided by the different analysis engines*" (ENSURE, 2012).

1.5 Aim and Objectives

"The aim of this research project is to develop a framework to predict the whole lifecycle cost of carrying long–term digital preservation in the cloud computing environment. The developed framework and its output cost model will focus on serving three business sectors; healthcare, finance and clinical trials."

The main objectives of this research project are to:

- Develop and validate a framework for long-term digital preservation cost
- Identify business sector requirements differences regarding their digital preservation needs
- Identify work and cost breakdown structures and the cost drivers associated with digital preservation
- Define an uncertainty identification process and incorporate uncertainty and obsolescence impact factors into cost prediction within proposed framework

1.6 Thesis Structure

The remainder of this thesis is composed of six chapters, as illustrated in Figure **1-3**. Chapter (2) reviews previous work in cost modelling in general, on digital preservation and on cost modelling for digital preservation. It also discusses cloud computing and its place in digital preservation as an example of storage and compute technology that is used now.

Chapter (3) summarises the research methodology of this project. A brief discussion of research methodologies in general follows, with specific reference to the rationale behind the decisions taken for this project. The results achieved lead to a single point estimate output for the designed framework presented in Chapter (4). It fully discusses the construction of an LTDP single-point cost model. In Chapter (5) the uncertainties and obsolescence issues found and their direct and indirect impact on costs is shown. An uncertainty identification process is discussed, along with a detailed obsolescence taxonomy claimed to present most of the obsolescence issues that could face any preservation practitioner. Following this the developed LTDP cost modelling process and framework are demonstrated.

The sixth chapter investigates the validation process of the framework through expert opinions and through a proof of concept tool. Diverse experts' experiences, different validation sessions and different questionnaires are cited to improve the accuracy of the validation process. Finally, conclusions are discussed, and the contribution to knowledge and future work are assessed in Chapter (7). This summarises the thesis results and shows what could be added to expand this area of knowledge.

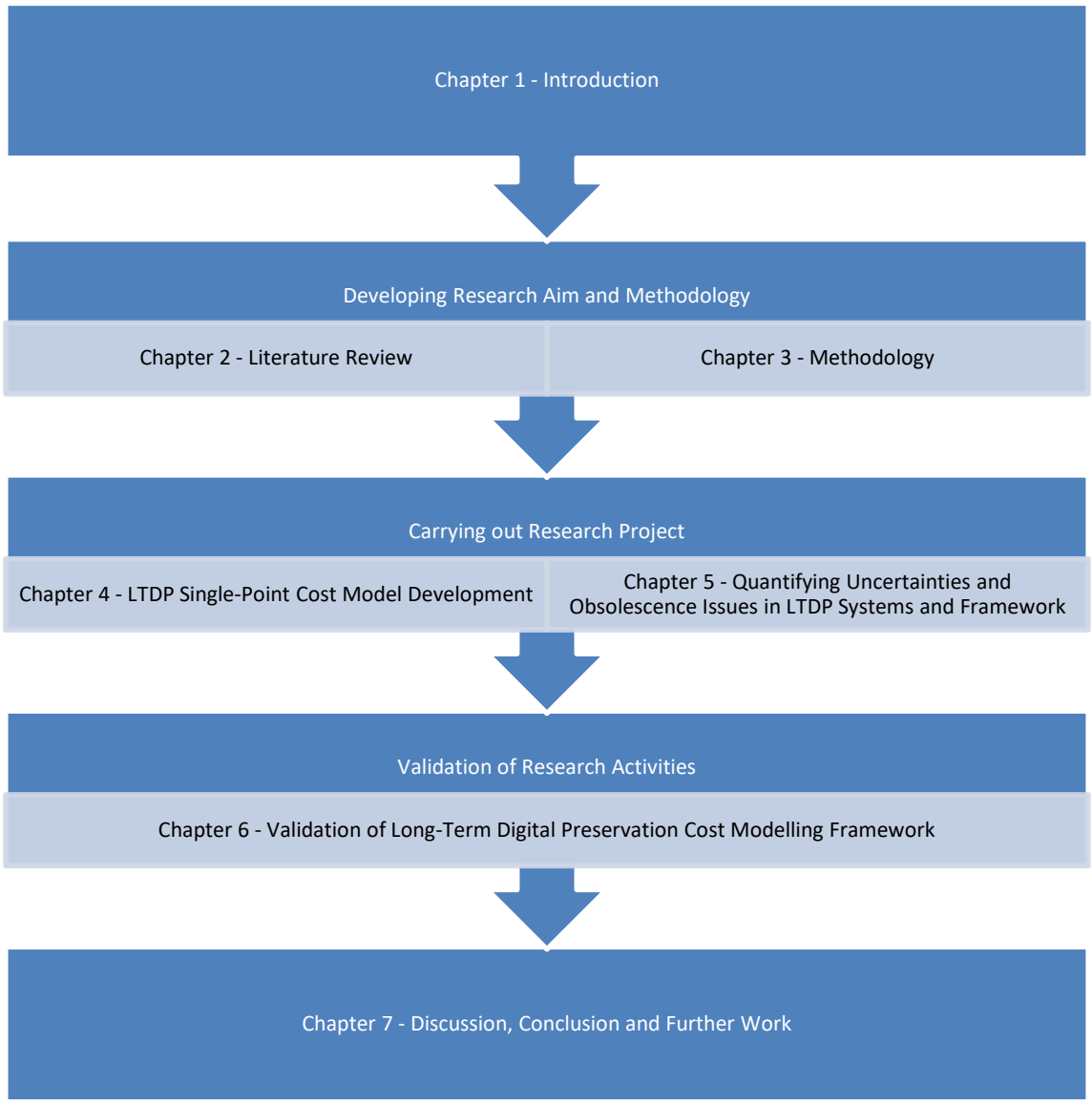


Figure 1-3 Thesis Chapters Structure

2 LITERATURE REVIEW

2.1 Introduction

Preserving information and ensuring its usability over long periods of time is a rising issue, with the amount of important data piling up and increasing these days in firms, government organisations and libraries. The importance of this information is a major driving force behind much of the development in this area. More and more information is changing through digitization from a material format to something digital form, but keeping this data usable for as long as an establishment requires can have economic implications. These implications derive from many factors, mainly precautionary activities to ensure the integrity, security and safety of the preserved information.

Building on these precautions, libraries and archives were the first to drive research into methods of planning and carrying out ways to preserve their continuously growing digital data, since they have traditionally served as the central institutional focus of preservation (Hedstrom, 1997; Corrado and Sandy, 2017). So many preservation initiatives were taken to help the libraries and archives that mainly faced this challenge that it was estimated by Hedstrom (1997), to be a time bomb.

These preservation initiatives generated standards and techniques and identified a need to estimate preservation costs. It was done through various cost models, designed to estimate the cost of preserving data over the long term, while considering infrastructure, the business understanding of digital preservation, compliance with any legal requirements, the long-term integrity and authenticity

of data and the use of commercially available IT technologies (Waddington et al 2016).

This chapter focuses on the main research areas of the research project, as shown in **Figure 2-1**, exploring work done previously in the fields of digital preservation and cost modelling techniques. These main sections are the core fundamentals required to proceed with developing a suitable cost model for long-term digital preservation using cloud computing technologies.

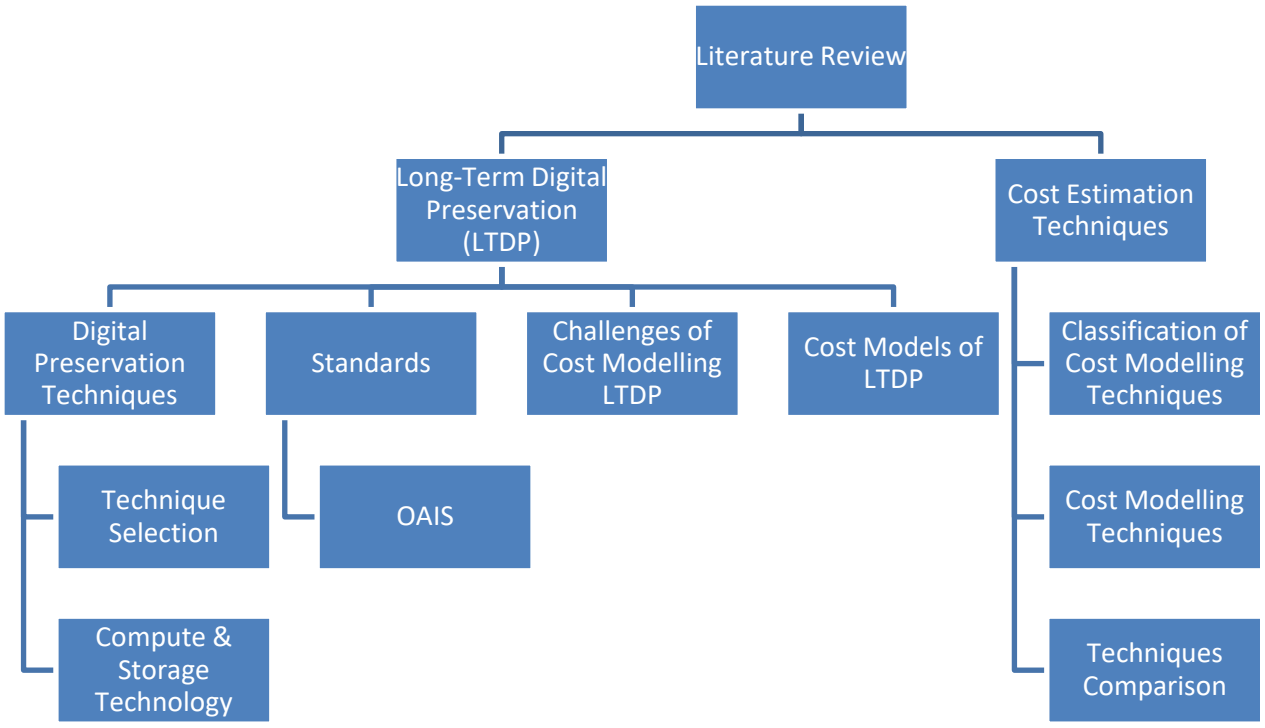


Figure 2-1 Structure of the Literature Review

2.2 Long–Term Digital Preservation

Introducing long–term digital preservation (LTDP) and its key areas, this section discusses LTDP techniques, standards, challenges and previously developed cost models.

Information nowadays is faced with many challenges regarding its future existence (Kuny, 1998; Hedstrom, 1997). The evolution of technology is so fast that recent information could easily be trapped behind obsolete hardware or software (Sandborn, 2007). Challenges, like those of obsolescence, generate warnings about preserving new datasets and information needs a different strategy, rather than simple storage actions. Storage should not involve merely keeping a pile of data, but should keep them understandable, meaningful, accessible, usable and useful over time as its main target.

The challenges facing digital information preservation have been discussed by many authors (Kuny, 1998; Hofman, 2009; Kay et al. 2014; Mayer et al. 2015; Waddington et al 2016; Corrado and Sandy, 2017). The rapid increase in the volume of digital information, the evolution of new technologies and issues of technology obsolescence, require constant active management of its content and the availability of preservation solutions is still fragmented. The preservation community has to face four main challenges, which highlight the constant threat facing every day's generated information and had driven the major interest in researching digital preservation.

Noting the severity of preservation issues, some have gone as far as comparing them to a time bomb (Hedstrom, 1997) or this era to the dark ages (Kuny, 1998).

Kuny (1998) and Hedstorm (1997) compare people's concern for centuries to preserve culture and history with their indifference nowadays to the vast amount of data lost through the mismanagement of information sources. In this section, we probe deeper into long-term digital preservation, exploring the techniques and standards available.

2.2.1 Long-Term Digital Preservation Techniques

Lee et al. (2002) discuss the evolution of data storage and the progressive development of digital preservation as shown in **Figure 2-2**. The argument shows the evolution of storage media versus life time compared to the evolution of digital preservation techniques and strategies. It indicates the phases and formats that information has moved through over time, also indicating the main four preservation techniques in operation today, with space for further new techniques to be added in the future.

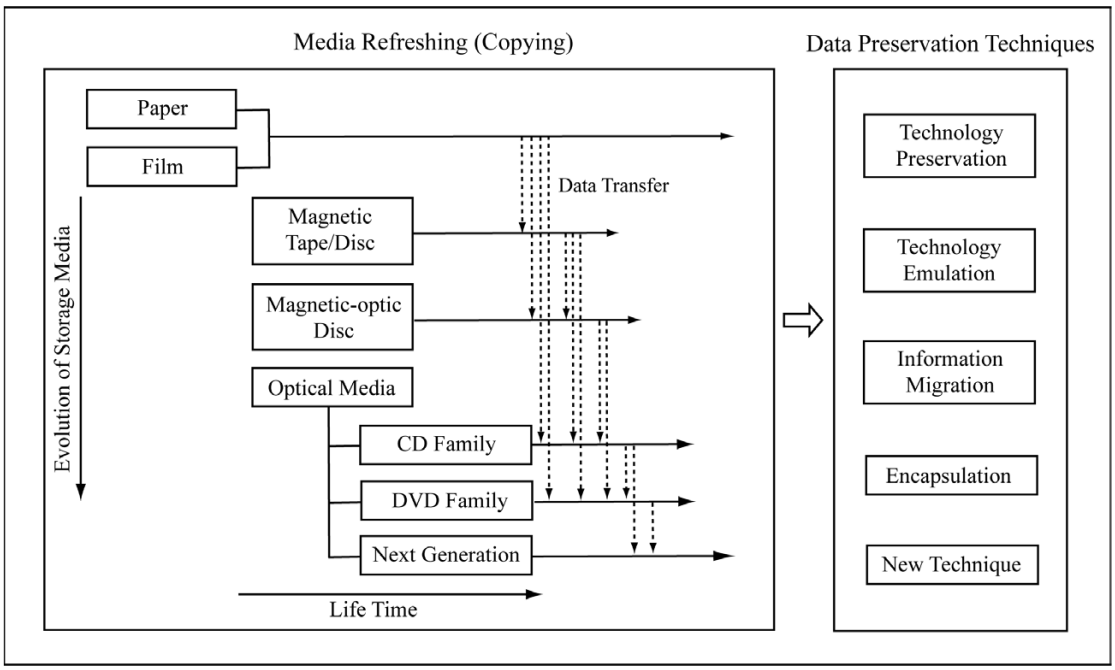


Figure 2-2 Evolution of Information Preservation Strategies (Lee, et al. 2002)

Authors have mentioned four techniques for long-term digital preservation (Corrado and Sandy, 2017; Borghoff, et al. 2006; Waugh, et al. 2000; Lee, et al. 2002; NLOA, 1999). Waugh et al. (2000) summarize the keys to successful long-term preservation.

Preservationists and data generators should first minimise the dependency of information on systems, other data or external documentation; which in turn will reduce the impact of technological obsolescence. This will enable future users to find or develop software that can extract useful information from preserved data and make documentation capable of holding decodable preserved information. These suggestions require wrapping information to preserve with its descriptive metadata in a single location and to make the preserved information reachable by the preserving organization. As shown in **Figure 2-3**, existing preservation techniques are designed either to aim to preserve technological environment or to overcome obsolescence issues from file formats.

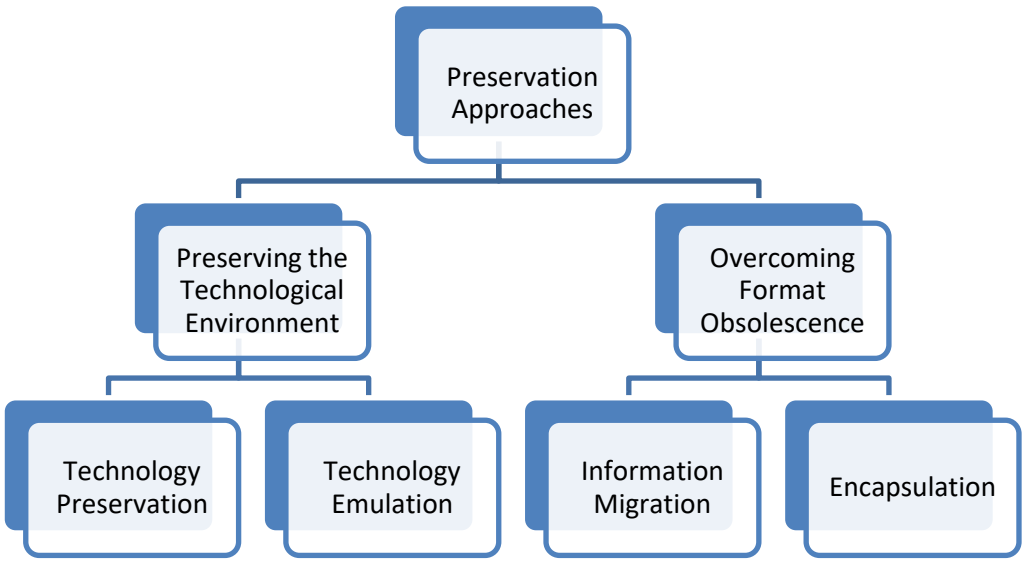


Figure 2-3 Preservation Techniques (Rothenberg, 1995; Waters and Garrett, 1996; Waught, et al. 2000)

2.2.2 Museum Approach (Technology Preservation)

Museum approach is meant to keep the original hardware, software, operating system, applications and everything that involves reaching the information as it currently stands. This preservation technique is best for the short term (Lee, et al. 2002) and will guarantee exactly the same behaviour over time (Russell, 2000). Unfortunately, it is very difficult to adapt it to long-term preservation since it becomes very expensive and needs much storage space for all the hardware that needs to be preserved (Graham , 1993).

Moreover, it is highly liable to fail, as agreed by Graham (1993) and Giaretta (2011), since it is exposed to failure in its electronic components due to ageing, dust, humidity and the loss of the human skill to use or maintain them.

It is clear that this technique is excellent but, due to its high cost and fragility, effective only for very short preservation periods and for information sensitive to any hardware or software change.

2.2.3 Emulation

This technique is designed to preserve the original programme, and then to give access by means of software, the *Emulator.* The emulator is designed to be capable of running on future hardware and operating system platforms (Rothenberg, 2000).

The emulator is usually developed when information needs to be retrieved. This ensures its compatibility with future platforms. The emulator will mimic the behaviour of the old system in the new one, and acts as an interface between the original programme and the new platform (Granger, 2000; Hendley, 1998).

It has proved successful in the gaming industry; old games can now run on any platform. **Figure 2-4** shows the emulation process. Borghoff et al. (2006) mention that two things must be preserved – the digital object with its metadata and a thorough document description of the home platform, given as inputs to emulators.

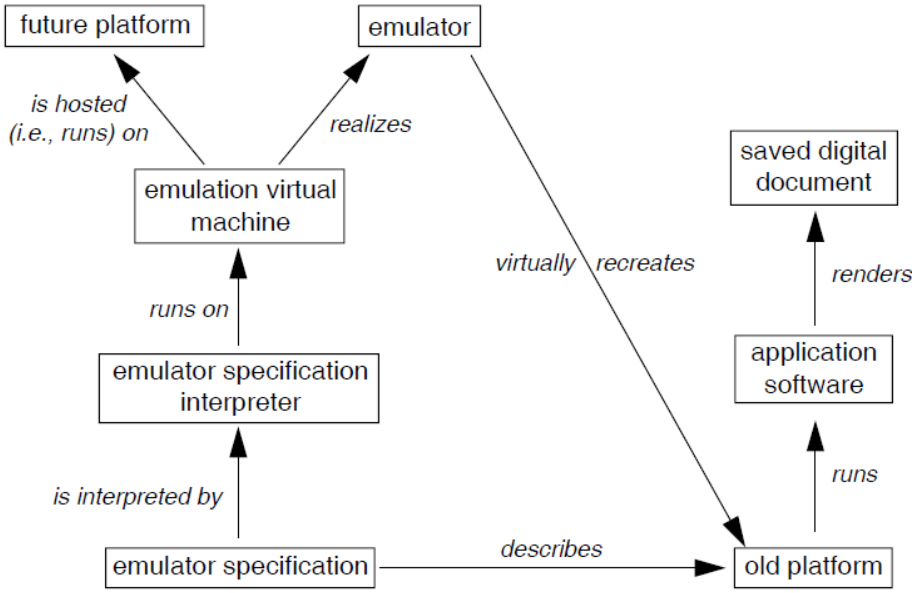


Figure 2-4 Emulation Preservation (Rothenberg, 2000)

Emulation is the most favoured technique for short to medium term preservation as seen by Hendley (1998); while it is favoured by Granger (2000) and Russell (2000) as most favoured for long-term preservation strategy, since it interfaces between the requirements of the information from the old system and what is available in the new system; this customisation is the strong point of emulation. Customisability leads to keeping information in its original context; no information is lost in moving formats (Rothenberg, 2000) and original documents stay with their original “binary stream” (Borghoff, et al. 2006).

While emulation as an approach promises well for the integrity and the look and feel of original information, its cost peaks at the start of documenting, because highly detailed and accurate documentation is essential, but also very complex to generate, especially when many formats are not yet standardised (Holdsworth, 2001; Holdsworth, 2006; Rothenberg, 1995). The amount of data to be kept alive has increased, from keeping both the information that needs preserving and documentation about the environment (Lee, et al. 2002), thus adding to the initial and running costs of the preservation system. Finally, it does not offer a solution to the obsolescence of human skill in handling preserved programmes and documentation (Corrado and Sandy, 2017; Waugh, et al. 2000).

2.2.4 Encapsulation

In encapsulation, the preserved records are wrapped inside a readable wrapper. This readable wrapper should include all the information needed to eliminate format obsolescence; **Figure 2-5** shows the structure of an encapsulated object.

Encapsulation can be combined with migration technique; all information will eventually need to migrate. However, a careful selection of starting and new formats and documentations will inhibit migration for a very long time (Day, 2006; Shepard, 1998). It is considered a passive technique and Lee et al. (2002) recommend it for data that will not be actively accessed. The recommendation based on this approach is simple, self-documenting, self-sufficient and can be combined successfully with the migration technique.

While Lee, et al. (2002) recommend the approach, they still warn of the difficulty of preserving file format information since it is not standardised for most formats and the methods for implementing the information in the wrap are vague.

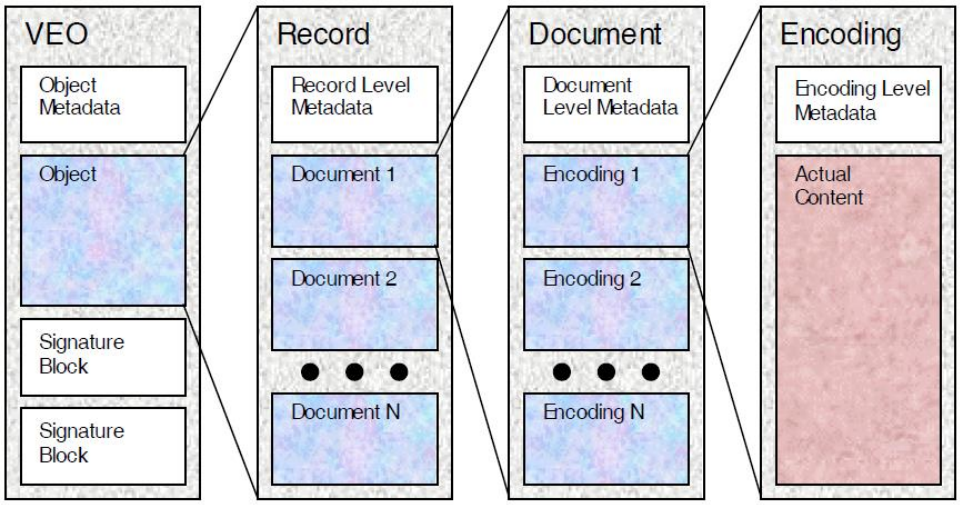


Figure 2-5 Structure of an Encapsulated Object (Waugh, et al. 2000)

2.2.5 Migration

The behavioural design of this technique is clear from its name; it denotes migration from an old, obsolescent and vulnerable technology to a newer, more stable one. Waters and Garrett (1996) define it as the periodic transfer of digital materials from one generation of hardware and/or software configuration to another later generation.

The Consultative Committee for Space Data Systems (CCSDS) recognizes three kinds of migration actions (CCSDS 2002):

1. Refreshment (ensures a reliable copy of the bit stream of the digital object)

2. Replication and Repackaging (ensures the availability of a manageable package).

3. Transformation (modifies the actual bit stream).

Migration is considered by many authors to be the most practical technique for short- and medium-term digital preservation (Russell, 2000), since no old technology or skill is to be preserved (Lee, 2002). Clear methods and tools are available because migration is a well-known technique, implemented in many I.T. departments (Borghoff, 2006). The most interesting benefit of migration is the instant availability of information, since it is already available in a current format and users are familiar with it (Rothenberg, 1999; Waugh, et al. 2000; Lee, 2002).

Russell (2000) argues that in the longer-term costs can increase more than expected over time, while Borghoff (2006) and Lawrence (et al. 2000) fear that it could become time-consuming if a large amount of digital material needed to be migrated, and if many different technologies were involved or if the information is on very diverse types of records and hardware. The degradation of the information's authenticity is an issue with migration; after a time, the original document is lost, thus reducing the authenticity and original character of the information. This could result in some data loss through a chain of migration activities and different technologies will require different migration strategies (Lawrence, et al. 2000).

2.2.6 Technique Selection

From the above discussion of preservation techniques, choosing a suitable preservation strategy is not simple. Each technique can benefit the system but is still limited by the nature of its design. Lee (2002) proposes a selection process, shown in **Figure 2-6**, which should make the choosing of a technique easier. It is

composed of a set of questions, regarding the complexity of the data, the knowledge about formats and the intent whether or not to actively access the data.

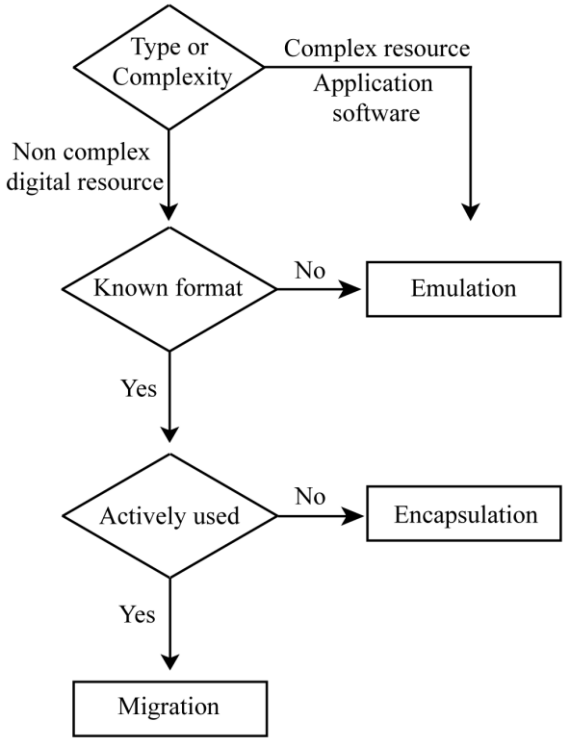


Figure 2-6 Selection of Preservation Technique (Lee, 2002)

2.2.7 Digital Preservation Standards

Like any procedure that companies and establishments commit to, digital preservation has some guiding standards.

1. Open Archival Information System (**OAIS**).

2. Preservation Metadata: Implementation Strategies (**PREMIS**) (Appendix E - E.1)

3. Network of Expertise in long-term STOrage of digital Resources (**NESTOR**). (Appendix E - E.2)

4. Digital Repository Audit Method Based on Risk Assessment (**DRAMBORA**). (Appendix E - E.3)

5. Trustworthy Repositories Audit & Certification: Criteria and Checklist (**TRAC**). (Appendix E - E.4)

These standards are the most commonly used by organisations attempting to preserve data. A detailed discussion of the OAIS reference model standard is the focus of this section. All other standards are discussed in the appendix of this document (Appendix E).

2.2.7.1 OAIS Reference Model

An ISO standard defined by the Consultative Committee for Space Data Systems (CCSDS), the model is defined in the CCSDS recommendation OAIS report as “An archive, consisting of an organization of people and systems, which has accepted the responsibility to preserve information and make it available for a Designated Community” (CCSDS, 2002). This defines the framework for a successful repository (Higgins, 2009; Corrado and Sandy, 2017).

A major purpose of this reference model is to facilitate a much wider understanding of what is required to preserve and access information for the long term (CCSDS, 2002). CCSDS defines the environment in which the archive functions as having three interfaces (see **Figure 2-7**).

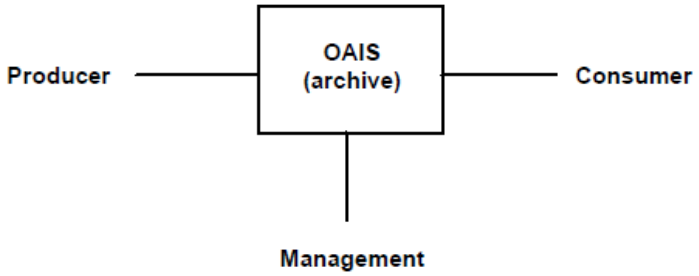


Figure 2-7 Environment Model of an OAIS (CCSDS, 2002)

Here a producer is the provider of the information to be preserved, management are those who set the overall policy and consumers are users who access the OAIS to retrieve preserved information.

Looking inside the OAIS box, **Figure 2-8** shows how it should function and the relationships between different entities in the suggested system. The OAIS model also describes each interaction between the entities and indicates how information packages will be handled in and between entities.

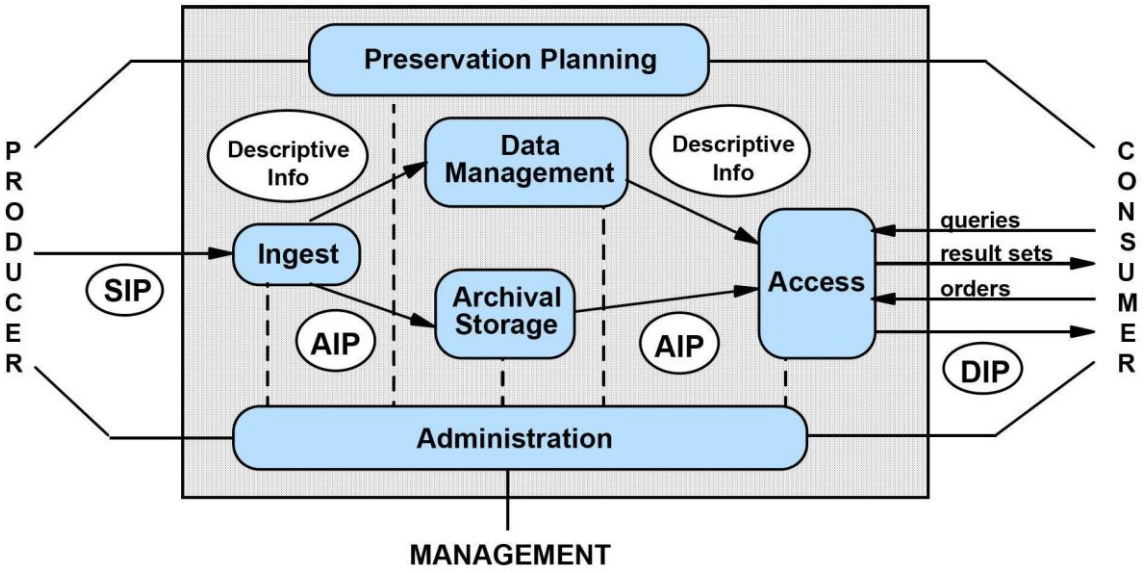


Figure 2-8 OAIS Functional Entities (CCSDS, 2002)

The diagram shows the integral entities/activities, which interact with each other bearing data packages and descriptive information. There are three types of

information package wrapped in descriptive information. The Submission Information Package (SIP), the data package input by the user to the OAIS, will have some content information and preservation description information (PDI). An Archival Information Package (AIP) is generated inside the OAIS from the SIP. It contains a complete set of PDIs.

The Dissemination Information Package (DIP) is provided to the consumer upon requesting an AIP, whole or in part. The Descriptive Information (DI) is the information used to identify packages inside the OAIS and is used to make these packages discoverable (CCSDS, 2002). Each information package is digitally constructed with its corresponding DI, as shown in **Figure 2-9**.

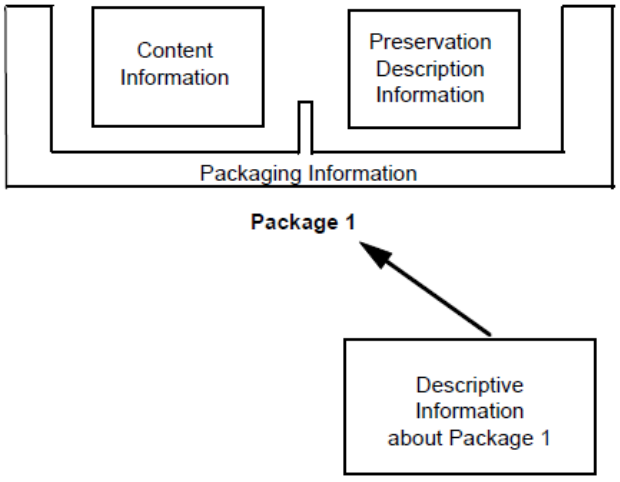


Figure 2-9 Concept of Information Package (CCSDS, 2002)

After defining each information package, the following are all activities that receive and/or generate information packages in the OAIS model. Ingest accepts SIPs from the producer and prepares the contents for storage and management in the OAIS. Archival Storage is where it receives the AIPs from ingestion and populates them into storage, manages the hierarchy, refreshes the medium to

avoid loss of data due to media failure, performs error checks, supplies recovery and provides AIPs when accessed.

What Data Management does is to populate, maintains and access the DIs and administrative data, which are used to manage the archive. It administers archive database functions, performs database updates, addresses queries on data management information to produce sets of results and creates reports from these sets. In Administration, the activity oversees the whole operation of the archive system. It audits SIPs and maintains the configuration of the system's hardware and software. It also monitors and improves archive operations.

In Preservation Planning, the activity monitors the OAIS environment and makes recommendations to ensure that the stored information is accessible to designated users for a long time. The Access activity provides consumer support, by looking up the requested information, acquiring its description, finding its location and availability and allowing the user to request and receive this information.

The OAIS reference model is considered by many to be the corner stone of modern strategic design for digital preservation. Most of its initiatives were designed to follow the OAIS; while some projects have chosen to strictly follow all the recommendations, others do not follow them in every detail.

This is due to the diversity of requirements between business sectors and between organisations. This helps each organisation when it begins to take preservation seriously to find the most suitable strategic design for its digital preservation system.

2.2.8 Current Compute and Storage Technology

This section discusses cloud computing technology and its cost structure. Cloud computing was chosen by the researcher to represent compute and storage technologies due to two main reasons:

- High similarities with previous technologies, e.g. Cluster computers and grid computers
- The current technology utilised by sophisticated digital assets owners, e.g. banks, social media companies and major technology providers
- It is not foreseen to be obsolete soon and known for its high upgradability

2.2.8.1 Cloud Computing

Cloud Computing as a concept has been imagined as vision since the early 1960s. This vision steadily evolved until the late 1990s, **Figure 2-10Error! Reference source not found.**, when it was realised in grid computing, which then evolved into cloud computing (Pallis, 2010). Mell and Grance (2011), of the National Institute of Standards and Technology (NIST) USA, defined Cloud Computing as “a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction”. The cloud computing model consists of three service models (see **Figure 2-11Error! Reference source not found.**), five essential characteristics and four deployment models (Foster, et al. 2008; Mell and Grance 2011).

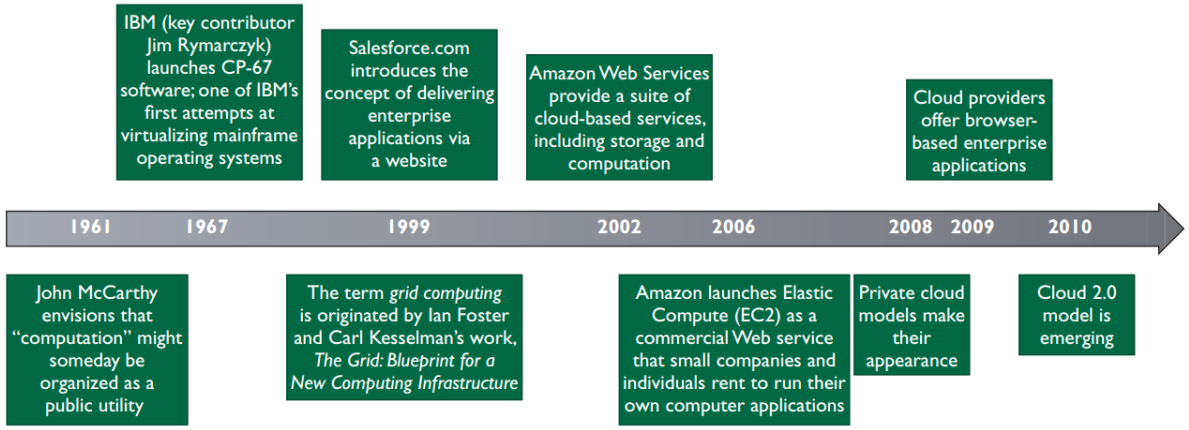


Figure 2-10 Timeline of cloud computing evolution (Pallis, 2010)

2.2.8.1.1 Service Models

Three main service models exist for cloud computing; Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). Having main models does not preclude the existence of other models. These are considered the main service models only because other models are designed around one or more of them (Foster, et al. 2008; Gong, et al. 2010; Mell and Grance 2011).

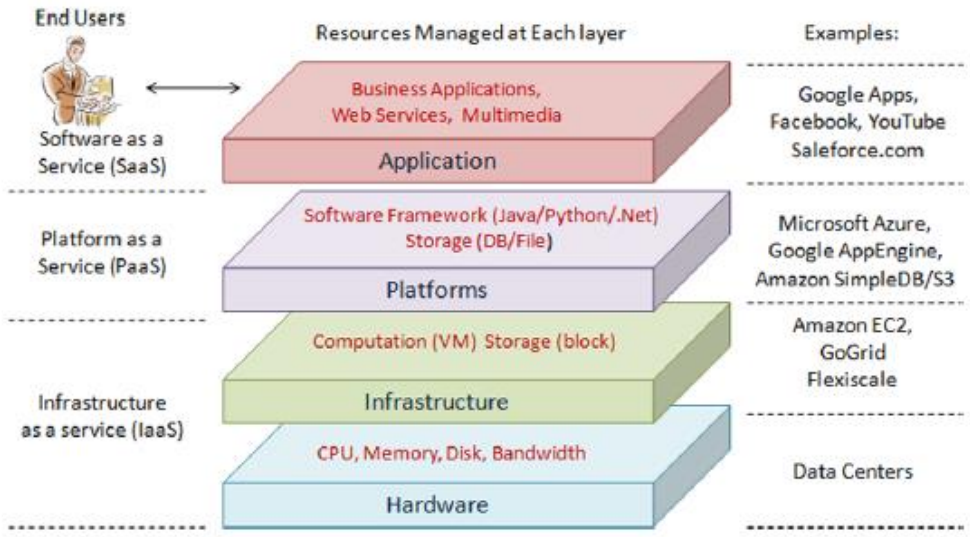


Figure 2-11 Cloud Computing Architecture (Zhang, et al. 2010)

SaaS is the capability of providing users with a software application that runs on cloud hardware and software infrastructure resources. PaaS is the capacity to enable users to deploy their generated or acquired applications, using cloud hardware and software infrastructure. The programming languages, libraries, services and tools of these applications must be supported by the cloud provider. IaaS is the ability to provide users with fundamental computing resources, i.e. provision processing, storage, networks, etc., where the consumer is able to run software, either operating systems or applications (Zhang, et al. 2010).

2.2.8.1.2 Characteristics

The nature of cloud computing is that it is an On-Demand Self Service, whose consumers automatically gain some provisional computing capabilities. It should also provide Broad Network Access; available and accessible capabilities over the network, through standard mechanisms, regardless of platform; e.g. pcs, tablets or mobile phones.

Resource Pooling: in cloud computing the resources of the provider are pooled to serve multiple users. It has high location independence, since users can hardly choose where these computations take place. Minimum location control is provided, i.e. country or data-centre. One of the beneficial core characteristics of cloud computing is its Rapid Elasticity. It has elastically provisioned and released capabilities to suit the scalability of demand. From the consumer's point of view, provisioning capabilities often appear to be unlimited. Measured Service, based on automatic control and resources optimisation in the use of cloud systems, is hired by metering capabilities. Resources are monitored and controlled (Mell and Grance 2011).

2.2.8.1.3 Deployment Models

There are four major cloud deployment models; private, public, community and hybrid computing clouds, as most authors agree (Pallis, 2010; Mell and Grance 2011; Zissis, et al. 2012).

Private Cloud is provisioned by a single organisation, serving multiple consumers, and it can exist on or off site. It can be managed, owned and operated by the organisation, a third party or both. Public Cloud is provisioned for open use by the general public. It will only exist on the provider's site and will be owned, managed and operated by a business, academic or governmental organisation.

Community Cloud is provisioned by an exclusive community of consumers with shared interests and can exist on or off site. It can be managed, owned and operated by one or more of the organisations in the community, a third party or both. Hybrid Cloud infrastructure is composed of two or more cloud infrastructures. They remain unique but are bound together through technologies which make data portable.

2.2.8.1.4 Private Cloud Computing and Data Centres Costs

Most of the initial cost benefits of cloud computing are realised for public cloud users, where the initial costs are very low and match what they require, on a pay-as-you-go basis. Sometime the data owners cannot export information outside the organisation and their only recourse is the private cloud. The cost structure is completely different as a breakdown of the essentials shows.

A private cloud cost realisation has five main elements (Greenberg, 2008) (see **Figure 2-12Error! Reference source not found.**): power management and

ventilation, use and latency optimisation, power consumption, servers and computational power and networking equipment. These individual cost elements are added to the setup and running costs of the infrastructure.

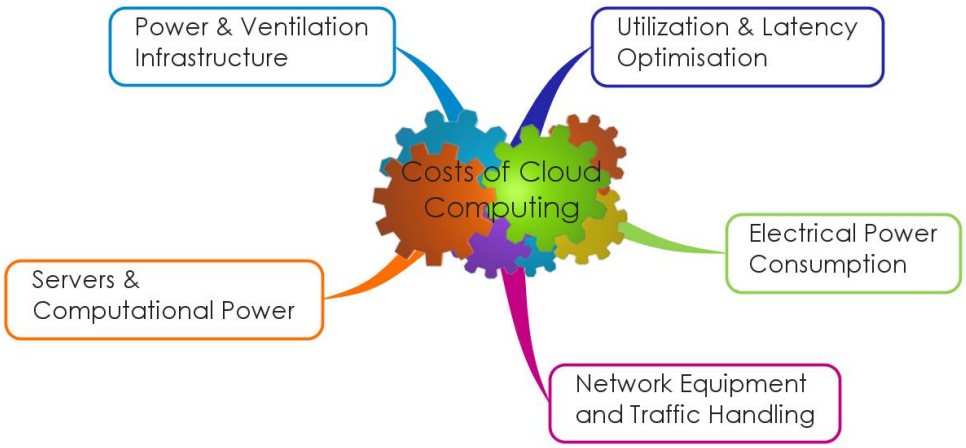


Figure 2-12 Main Cost Areas for the Setup and Running of a Private Cloud

The cost of power and ventilation infrastructure is the cost of delivering consistent power and evacuating heat. Most of the cost is a one-off payment whose high value reflects the demand for this type and level of servers' concentration; while some of it, for the maintenance of the system, is on-going. Servers & computational power adds to the cost of purchasing proper computational power, to achieve high use, which must guarantee the pooling of proper resources. Electrical Power Consumption costs cover the electricity drawn by the servers and heat evacuation systems. These are on-going costs; the servers' concentration will determine the rate of growth of these costs.

Network equipment and traffic handling are the capital elements of the networking cost and are mostly concentrated in the networking gear. The remainder is divided between traffic handling linking the internet service providers and the end

user, inter-cloud links between different geographical locations and finally the regional facilities needed to reach wide area network interconnection sites. The cost of use and latency optimisation is to avoid low use of one's cloud facilities and avoid low latency to end users

2.3 Challenges of Cost Modelling for Long-Term Digital Preservation

Developing a cost model is a task that requires information, rigorous research and a good understanding of the expected challenges. These challenges differ from one model to another and from one business sector to another. In long-term digital preservation, the challenges are found in four main areas, as shown in **Figure 2-13** (Xue, et al. 2011): Technological, Information, Methodological and Business- oriented challenges.

Technological challenges are generated from the nature of technology itself, its obsolescence and the uncertainty surrounding existing file formats (Romero Rojo, 2011; Erkoyuncu, et. al, 2009). Other technological challenges have come from understanding the cost of migrating information (Russell, 2000) and the cost of technologies used in new LTDP projects, such as cloud computing (Rosenthal, et al. 2010; Baker, et al. 2006).

Cost Information poses a great challenge since it is not easily available; collecting cost data is not easy (Roy, et al. 2001), different data formats generate different costs and the existence of uncertainties skews the estimations.

The challenges of methodology are, first, to learn how a cost model can be internally evaluated and to find the cost details for digital preservation; but these

are still not enough. It is also necessary to find how technical recommendations can be generated from cost model and from generic cost estimation techniques. Predicting LTDP costs for new business ventures will always generate unknown challenges in each business sector.

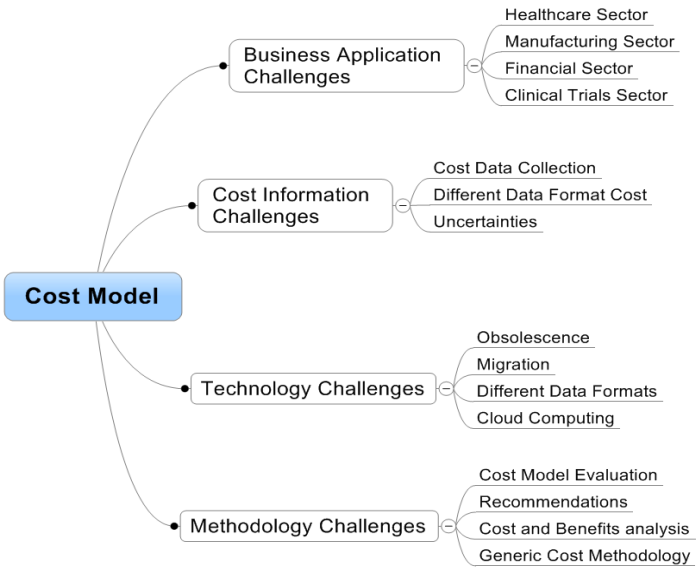


Figure 2-13 Cost Modelling Challenges for LTDP (Xue, et al. 2011)

2.4 Cost Modelling for Digital Preservation

Many cost models have been developed or are in process of development. In this section, these cost models will be reviewed and analysed. The main targets that these projects serve are either heritage or scientific data concerns. The following figure (**Figure 2-14**) shows the various significant cost models and the sector that they serve.

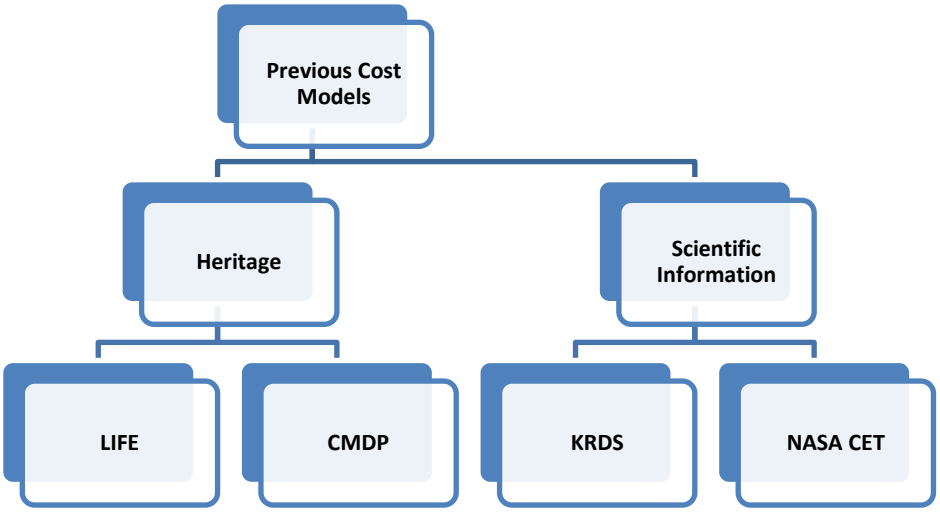


Figure 2-14 Cost Models for Digital Preservation

2.4.1 Lifecycle Information for E-literature (LIFE)

There are three stages to cost modelling project, LIFE[1], LIFE[2] and LIFE[3]. All three stages are funded by its main collaborators, namely, University College London (UCL) and the British Library (LIFE, 2007). All three stages together generate an activity based cost (ABC) model (Wheatley, et al 2007; Ayris, et al 2008; Kejser, 2009; Hole, et al 2010).

2.4.1.1 LIFE[1]

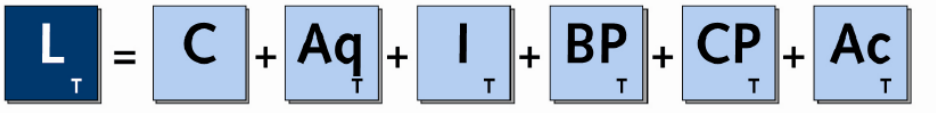
LIFE[1] is a one-year project aiming to explore the lifecycle approach in costing digital preservation. In 2007, LIFE[1] developed a generic model of a digital preservation lifecycle. It used three case studies (e-journals, web-archive and e-publications). **Figure 2-15** shows the main basic equation for LIFE[1] and its cost indicators. The estimator needs to fill these requirements to discover the total cost (L_T).

Acquisition	Ingest	Metadata	Access	Storage	Preservation
Selection	Quality Assurance	Characterisation	Reference Linking	Bit-stream Storage Costs	Technology Watch
IPR	Deposit	Descriptive	User Support		Preservation Tool Cost
Licensing	Holdings Update	Administrative	Access Mechanism		Preservation Metadata
Ordering & Invoicing					Preservation Action
Obtaining					Quality Assurance
Check-in					

Figure 2-15 LIFE[1] Cost Model and Cost Indicators (Wheatley, et al. 2007)

2.4.1.2 LIFE[2]

This research was carried out by the same funding bodies and research team of LIFE[1] working for one-and-a-half years with the aim of validating and refining the model developed in LIFE[1] (LIFE, 2008) and of investigating LIFE[1] by economic experts. Two case studies, SHERPA-LEAP and SHERPA-DP, both based on repositories development (Ayris, et al. 2008; LIFE, 2008) were made. The researchers used the Burney Collection of the British Library, a collection of newspapers and news pamphlets from the 17th – 18th century compiled by the Reverend Charles Burney (British Library, 2012), to help in "*enabling effective planning and decision making for the cost of preservation for collections that exist as both analogue and digital*" (LIFE, 2008). Finally, they reported lifecycle and preservation costing for analogue materials, material in repositories, primary data and digital surrogates (see **Figure 2-16**) (LIFE, 2008).



Lifecycle Stage	Creation or Purchase [43]	Acquisition	Ingest	Bit-stream Preservation	Content Preservation	Access
Lifecycle Elements	....	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 2-16 LIFE^2 Cost Model and Cost Indicators (Ayris, et al. 2008)

2.4.1.3 LIFE^3

This was a one-year project aiming to develop ways of estimating the cost of the digital preservation life cycle, by moving the focus of LIFE work from post-event analysis into predictive costing (Wheatley, et al. and Hole, et al. 2009). **Figure 2-17** shows the development of the predictive cost modelling tool for estimating key areas of preservation that are difficult to predict due to the lack of historical data (LIFE, 2010).

The target here is to give support in enhancing planning and decision-making activities (Wheatley, et al. and Hole, et al. 2009) with a simple cost modelling tool; it has a web tool interface which integrates other commonly used preservation standard models into the estimation tool (see **Figure 2-18**) (Hole, et al. 2009). LIFE^3 presents a cost model for each stage of the preservation lifecycle.

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				

Figure 2-17 LIFE^3 Cost Model and Cost Indicators (Hole, et al. 2010)

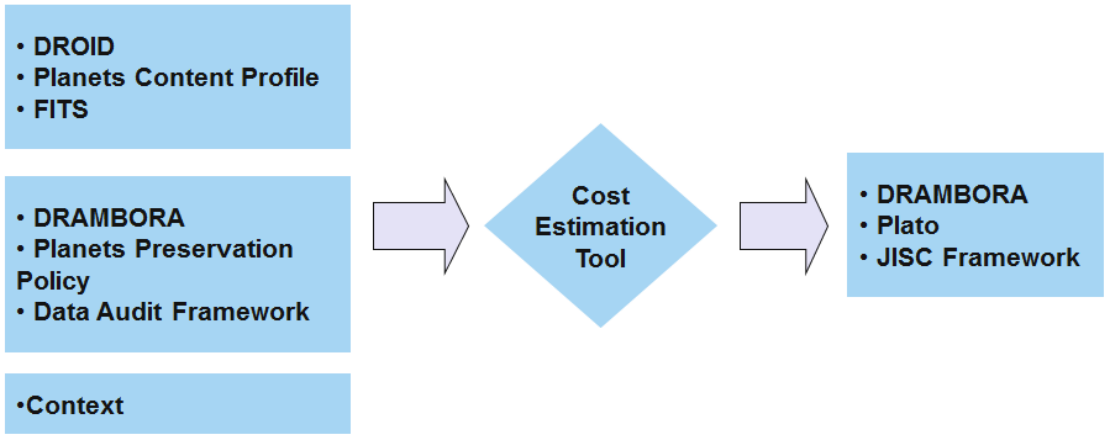


Figure 2-18 LIFE^3 Cost Model Integration with Standards and Tools (Hole, et al. 2010)

LIFE Analysis

LIFE in general has adopted a lifecycle approach to the cost estimation of digital preservation. It adds an extra step to the OAIS model, to enable preservation experts to create or purchase the data to preserve. This will reflect any cost dependency.

As a core collaborator, the British Library provided strong technical backing to LIFE. LIFE provides a more flexible version of the OAIS, since the acquisition stage which occurs before Ingest allows the organisation to choose the appropriate preserved information, and not save all its data. LIFE^3 has some similarity to CMDP in that it developed a detailed process level calculation.

Unfortunately, it is less user-oriented in design and has not been validated against actual cost values; it can be misleading. LIFE is generic but only in the heritage and library domain and it offers no real linkage between the stages of costing. Each is calculated in isolation, which can generate misleading values, since a choice at one stage may affect the costing in a subsequent stage (Kejser, 2009)

2.4.2 NASA Cost Estimation Toolkit (CET)

This cost model is specifically built to serve the scientific data centres of NASA, the National Aeronautics and Space Administration, and enable it to cost estimate the lifecycle cost of its data systems. The tool can run on PC or MAC, and is based on Excel Visual Basic for Applications (Ball, 2008; NASA, 2008)

The CET relies on the Comparables Database (CDB) containing historical information about 29 projects. It provides outputs in spreadsheet and graphic formats, and includes tools for what-If options, reviewing the output and manual override outputs, with sensitivity tests for parameters. Finally, it is a tool for adding new historical data for new projects (NASA, 2011). The current version for this toolkit is 2.4 and it can be downloaded from the toolkit's website.

The tool has a lifecycle approach to the cost of digital preservation, it uses a regression analysis cost estimation technique and looks at total cost as the sum of the staff effort and non-staff costs (Hunolt, 2008). **Table 2-1** shows the CET effort as a function of the workload equations.

Table 2-1 CET Effort as f(Workload) Relationships (Hunlot, 2008)

Linear	Y = a + b*X
Logarithmic	Y = a + b*lnX (ln is natural logarithm)
Exponential	Y = a*e^(b*X) (e is the base of the natural logarithms)
Quadratic	Y = a + b*X + c*X^2
Square Root	Y = a + b*X + c*sqrt(x) (sqrt - square root)
Linear-Logarithmic	Y = a + b*X + ln(X) (ln is the natural logarithm)
Linear-Exponential	Y = a + b*X + c*e^X (e is the base of the natural logarithms)

CET converts all categories of labour costs over the lifecycle of the data activity, into Full Time Equivalents (FTEs). Y is dependent variable (effort) and X is independent variable (workload) with a, b and c are coefficients computed by regression (Nasa, 2008). In linear, logarithmic and exponential relationships CET uses a "*single parameter regression of Y's on X's*" and in quadratic, square root, linear-logarithmic and linear-exponential it uses "*two parameter multiple regression*" (Hunlot, 2008). Regression analysis is discussed in section 2.6.1.7

The conversion from effort into staff costs is made by applying labour rates and the inflation rate to the effort estimates; after this the tool adds the non-staff costs, e.g. infrastructure, computers, etc. (Hunolt, 2008).

NASA's CET Analysis:

Similar to LIFE, NASA's CET also adopts a lifecycle approach to costs and is the most developed cost model with the highest number of available documentations. It is based on information from 29 projects, which defines the experience of the toolkit. The metadata have built-in fields that are used as key cost variables, which can be used to inter-link the stages of digital preservation with the components of the cost model. This also permits the sensitivity tests. NASA estimates that the error in the toolkit's estimation is at 22.9%, which reflects its strength in validating the output.

But it is very expensive to develop this cost model; it costs NASA $250,000 to $350,000 per year to maintain the database and expand the model and it does not function to estimate the cost of long-term digital preservation, but is limited to the calculation of current costs. It is also limited to serving only NASA's space and earth observation research.

2.4.3 Keeping Research Data Safe (KRDS)

The KRDS (2008-2010) cost model is funded by the Joint Information Systems Committee (JISC), UK and was developed by Charles Beagrie. He developed the cost model in two stages, KRDS1 and KRDS2, and identified the cost variables for preserving research data in UK universities. The model was designed on the basis of the OAIS reference model, LIFE projects and NASA CET and used 4

case studies (Beagrie, et al. 2008; Beagrie, et al. 2010). The target was to "*identify and analyse sources of long-lived data and develop longitudinal data on associated preservation costs and benefits*" (Beagrie, et al. 2008).

KRDS Analysis:

KRDS was developed and built on OAIS, LIFE and NASA CET, taking a Lifecycle costs approach. It was integrated with the TRAC auditing standard, commonly used in UK universities (Kejser, 2009). It was validated against real cost data from UK universities and it introduces the concept of economic benefits.

However, KRDS is designed for research data only; it does not strictly follow the OAIS reference model. It has many generic features but lacks specificity.

2.4.4 Cost Model for Digital Preservation (CMDP)

Also known as the Cost Model for Digital Curation (CMDC), this cost model was initiated in 2009-2011 by the Danish Royal Library, along with the State and University Library and the Danish National Archives, to estimate costs for digital preservation. It uses the OAIS reference model along with activity based costing (ABC) (Kejser, et al. 2009), to estimate preservation costs. CMDP divides the OAIS functions into delineated cost critical activities. To reach the total cost, the model sees Ingest and migration costs in detail, and then adds to them the cost of archiving. CMDP is divided into two phases.

The target of phase 1 (CMDP 1) was to populate the costs of logical preservation, based on the preservation strategies with migration activities only (Kejser, et al. 2009). **Figure 2-19** shows the main structure of the cost model; this is simply a high-level view. The model employs a more detailed approach to cost elements.

It was followed by phase 2 (CMDP 2) which focused mainly on the ingestion costs to the preservation system. Both phases employed an ABC estimation technique, and relied heavily on the OAIS reference model for the preservation activities (CMDP 2, 2010).

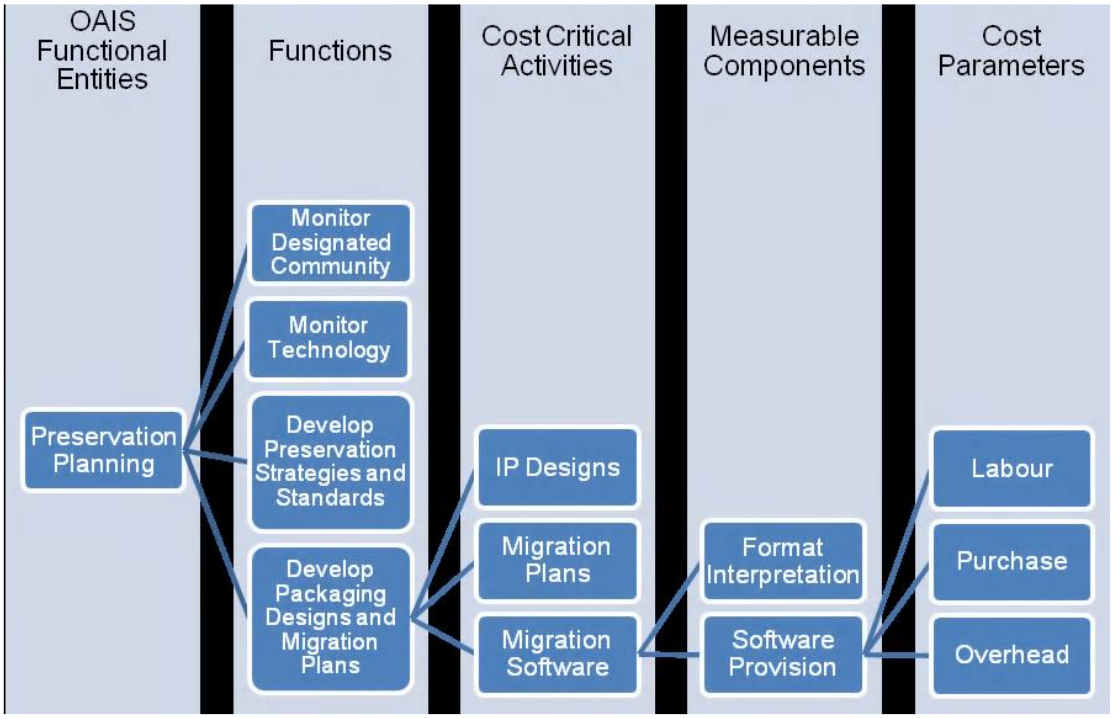


Figure 2-19 CMDP Structure (Kejser, et al. 2009; Kejser, 2009)

In CMDP 2 higher costs were estimated if the preserved data were emulated over time, rather than taking the cheaper option of migration, as shown in **Figure 2-20**. This is due to the initial excessive costs of algorithm extraction. This emulation cost drops over a longer period, but is still higher than the migration cost, due to the availability of the extracted algorithm.

KRDS is basically designed to follow strictly the OAIS reference model. It has a detailed Ingest and migration cost analysis and is validated against values from the Danish Library. Model was packaged and deployed as an Excel based tool.

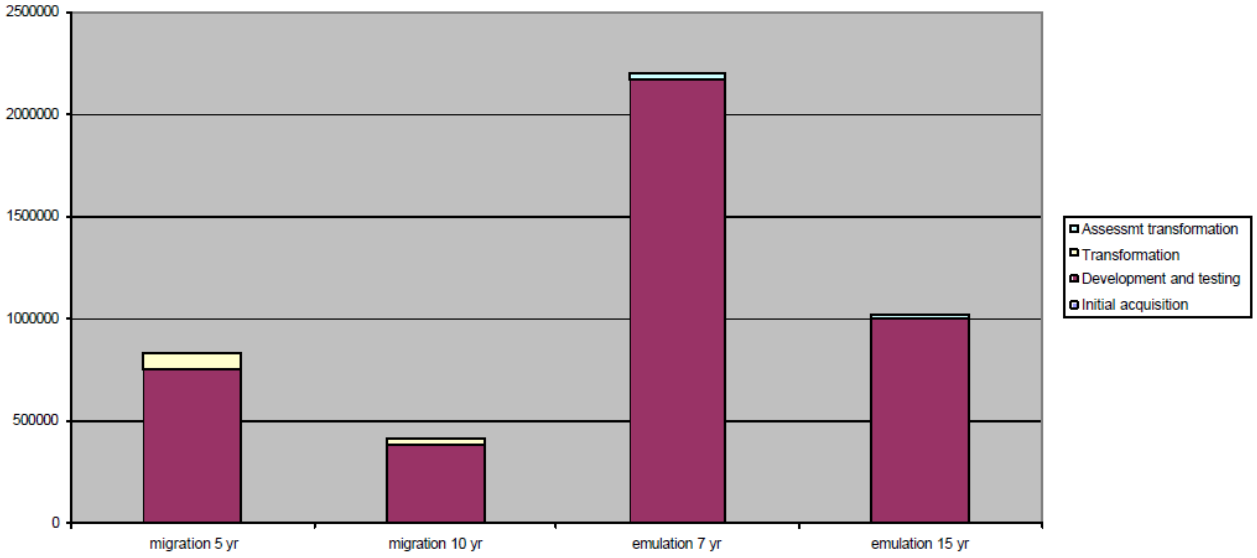


Figure 2-20 Cost Difference - Emulation and Migration (15 Years/€) (CMDP 2, 2011)

KRDS was designed solely for a specific sector, that of research and heritage. Thus, applying it in other cases for different business sectors requires extra work and careful implementation, since it too also does not apply a lifecycle approach. This is still a work in progress and still expanding.

.

2.5 Cost Estimation and Modelling Techniques

Cost, as explained by Evans (2005), is the "expenditure necessary for the attainment of a goal". Roy (2003) and Shehab and Abdalla (2001) define cost estimating as "*concerned with the predication of costs related to a set of activities before they have actually been executed*".

The Association for Advancement of Cost Engineering International (ACCEI) in 2007 issued the final version of its Recommended Practice Standard No. 10S-90, containing cost engineering terms. ACCEI defined cost as cash expenditure or liability incurred in considering goods and/or services. Cost may include the investment of resources in strategic assets. ACCEI also defines cost estimation as "*A prediction of quantities, cost, and/or price of resources required by the scope of an asset investment option, activity, or project, as a prediction, an estimate must address risks and uncertainties."* (AACEI, 2007)

Cost estimation techniques are used in industry to help companies improve their performance and oversee their spending. An accurate estimate is crucial to a product, as mentioned in many research papers (Roy, 2003; Niazi, et al 2006; Evans, 2005). Underestimation will lead to committing less resources to the work required, resulting in inability to complete it. In contrast, over-estimation could result in loss of competitiveness and the commitment of funds that will not be needed and that might leave other work commitments deprived.

This section of the chapter will concern cost modelling and the key areas of research related to it. It will discuss what cost estimation is, how to classify cost

modelling techniques and what the successful and proven techniques of cost modelling are. The structure of this section can be seen in **Figure 2-21**.

Figure 2-21 Structure of the Cost Modelling section

Roy in his paper (2003), as agreed by other authors (Niazi, et al 2006), argues the crucial importance of early cost estimation. This means that only early cost estimation can prepare companies for the spending that will be incurred by committing themselves to a certain project. This will give each company the strength to make more successful decisions over time. The reliability of the earliest estimates is not high, but their job is to give an indication of the size of the costs (Roy, 2003). This is needed for many reasons, mainly to do with the lack of data at such an early stage and the unpredictability of obsolescence issues over time. Selecting the right appropriate estimating technique is the decisive factor that will result in a high reliability cost model.

Cost estimation should not be confused with cost accounting; cost accounting focuses on the past consumption of resources, while cost estimation predicts future costs (Torp and Klakegg, 2016). Cost estimation techniques are used in cost models. Cost models are "*algorithms intended to replicate the cost performance of a process of a system*" (ACCEI, 2007).

2.5.1 Classification of Cost Modelling techniques

Many classifications seek to differentiate between different cost modelling techniques, required owing to the numerous techniques available and the desire

to make choosing between them easier. The two main classifications of cost modelling techniques are presented by Roy (2003) and Niazi, et al. (2006).

Roy (2003) devised a classification of cost models into five main categories (see **Figure 2-22**). This classification is somewhat simple and straightforward.

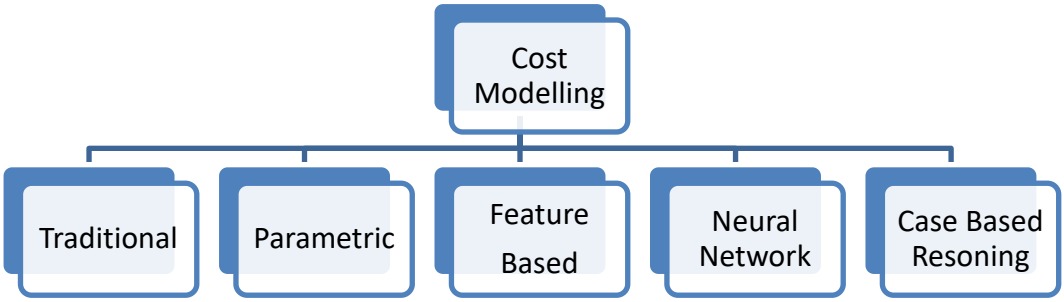


Figure 2-22: Cost Modelling Techniques Classification (Roy, 2003)

Other authors (Niazi, et al. 2006; Evans, 2005; Duverlie, 1999) agree on a somewhat different classification of cost modelling techniques. This classification is more complex in structure, but its depth confers a better understanding of what each technique and method can do. Niazi, et al. (2006) breaks the techniques down into two main categories; Qualitative (**Figure 2-23**) and Quantitative (**Figure 2-24**).

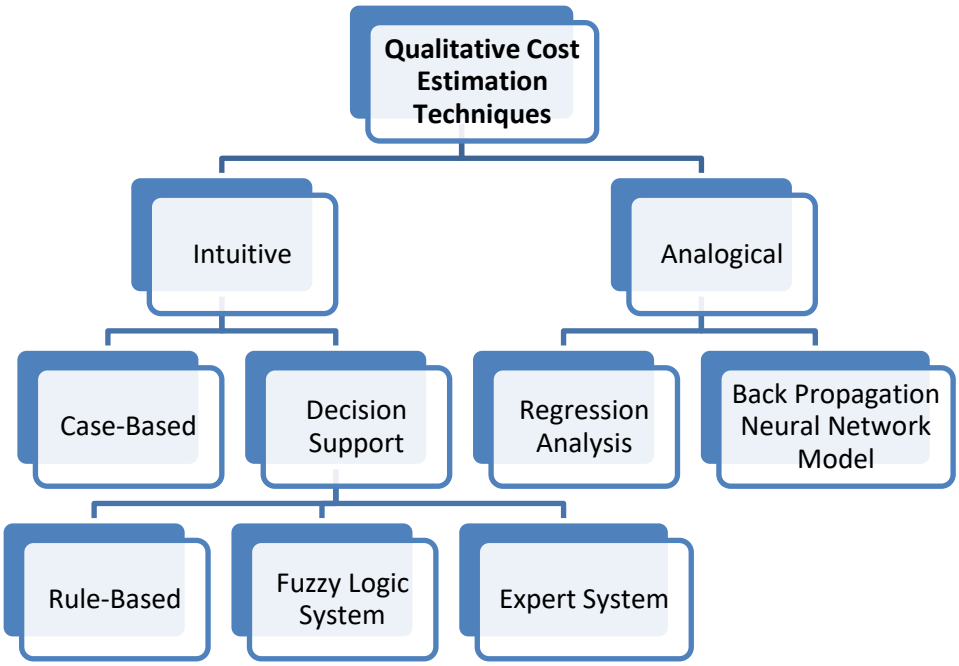


Figure 2-23 Classification for Qualitative Cost Modelling (Niazi, et al. 2006)

Qualitative cost estimation techniques are based on comparing new and old products in order to find similarities between them (Niazi, et al. 2006). Two cost estimation sub-techniques are considered qualitative, intuitive and analogical. The intuitive techniques employ an estimator's experience, based on his previous work (Niazi, et al. 2006); while analogical techniques find similarities between old and new products and base the costing on historical data (Niazi, et al. 2006).

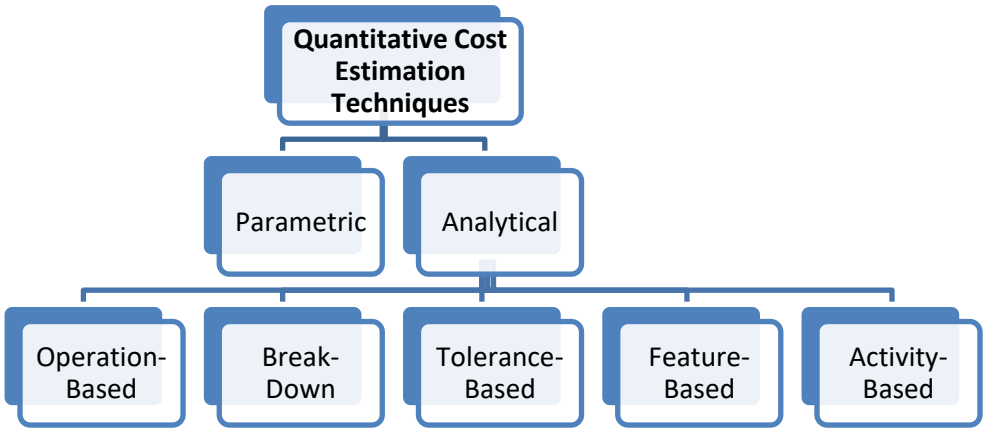


Figure 2-24 Classification for Quantitative Cost Modelling (Niazi, et al. 2006)

Quantitative cost estimation is based on analysing the detailed features, design and activities related to a product. Again, there are two current quantitative techniques, parametric and analytical techniques. The parametric cost estimation technique "*… uses Cost Estimating Relationships (CERs) and associated mathematical algorithms to establish cost estimates*" (NASA 1996), while analytical techniques require a product to be broken down into its basic elements, corresponding to all the resources used in the production cycle (Niazi, et al. 2006).

These classifications do not mean that a project has to follow a single cost model throughout its lifecycle. Duverlie (1999) advises that a project should combine several types of costing technique from the pre-design or feasibility phase to the services phase.

2.6 Cost Modelling Techniques

Cost estimation techniques build cost models. To be able to adopt a functional cost model, a thorough discussion is needed, which explores all the possible cost estimation techniques. From Niazi et al.'s (2006) classification, the breakdown of the techniques will be explored and the benefits and limitations of each technique will be highlighted.

Starting with qualitative techniques, two intuitive techniques are available: Case Based Reasoning (CBR) and Decision Based Systems (DBS). CBR uses the data of old designs and tries to adapt them to new design information, matching the attributes of old and new (Rehman, 1998). In DBS, expert estimators use these

models to make better judgements through the stored knowledge of field experts (Kingsman, 1997; Shehab, 2002).

In CBR techniques, Regression Analysis uses historical cost data to generate a linear relation between the costs of old and new products (Niazi, et al. 2006). Neural Networks use trainable neural networks, which can provide answers to unknown questions, based on the stored information that was used to train them (Edalew, et al. 2001).

DBS as an estimation technique has three sub-techniques: The Rule Based System (RBS), Fuzzy Logic System (FLS) and Expert System (ES). The RBS is based on the time and cost of feasible processes from a set of stored knowledge, incorporating production constraints (Gayretli and Abdalla, 1999). FLS is carried out by Fuzzy rules from a decision table, containing system rules and relations between inputs and outputs, which are used to handle uncertainty in cost estimation (Niazi, et al. 2006). Finally, ES is usually rule-based programming, used to mimic human logical reasoning, retrieving experience from databases (Niazi, et al. 2006; Venkatachalam, et al. 1993).

Among the quantitative techniques, discussed also on p. 65, above, two sub-techniques are Parametric and Analytic cost estimation techniques. The former is a standalone technique, but the five analytical sub-techniques are operation based, breakdown, tolerance-based, feature based and activity based costing.

The operation based approach estimates cost according to the total time needed to perform the tasks, corresponding to product manufacturing (Jung, 2002). The breakdown approach is usually employed at the end of manufacturing, due to the

number of details required. It sums up all the costs of manufacturing the product, from materials to overhead costs (Niazi, et al. 2006). A tolerance-based cost estimation takes into consideration the product design tolerances as a function of cost (Niazi, et al. 2006).

Feature-based cost estimation is concerned with identifying the cost-related features of a product; these contribute to the total product cost (Niazi, et al. 2006). Activity-based cost estimation calculates the cost by summing the cost of all the activities that are required for manufacturing a product (Andrade, et al. 1997).

After the above overview of the modelling techniques classified in Niazi's (2006) research, some popularly used models are explored in greater depth.

2.6.1.1 Case-Based Reasoning (CBR)

This technique uses solutions from previous experiences to resolve current issues (Duverlie, 1999). Case-based reasoning employs the databases of previous similar products that have similar cost data. Data from databases are used to find helpful information that it can match to new product characteristics (Roy, 2003; Niazi, et al. 2006). This technique functions in the following sequence of steps (Duverlie, 1999):

1. Recognition of problem.
2. Recall of similar experiences and their solutions.
3. Choosing and adapting a solution to the new problem.
4. Evaluating the new situation.
5. Learning from the solved problem.

This technique has an innovative design approach (Niazi, et al. 2006) which can find solutions rapidly (Duverlie, 1999). CBR has a transparent functionality, unlike other “black-box” techniques (Duverlie, 1999), which makes amending models based on CBR easy and future-proof, with a known solution source (Duverlie, 1999). Ultimately it preserves knowledge and does not depend on employees (Duverlie, 1999).

All these advantages come, however, with some limitations to the technique. CBR’s dependence on past cases is high (Niazi, et al. 2006), and these cases must have been validated (Evans, 2005). It cannot function with innovative products (Evans, 2005).

2.6.1.2 Fuzzy Logic

Fuzzy rules from a decision table, which contain system rules and relations between inputs and outputs, are used to handle uncertainty in cost estimation (Niazi, et al. 2006; Shehab, 2002). Fuzzy logic is developed in three steps: “*Fuzzification of inputs, fuzzy inference based on a defined set of rules and finally Defuzzification of the inferred fuzzy values. The main process in the fuzzy model is to assign fuzzy sets of input variables and fuzzy sets of output variables* “(Shehab, 2002).

Fuzzy logic has proved reliable in estimates even with uncertainties, but the price is tedium if estimating complex features (Niazi, et al. 2006).

2.6.1.3 Traditional Cost Estimation/ Expert Judgement (TC/EJ)

Rush and Roy (2001) discuss this technique. They define it as cost estimates generated by predictions from experts with enough years of experience and skills

in the specific field. Another definition (Evans, 2005) is that it is a process where humans from the domain in which the estimate is being made provide an estimate of the cost.

Though this has been widely used hitherto, Rush and Roy (2001) mention that it is not considered to be a cost estimating technique. But it can still provide a very quick estimate with little resource cost in terms of its requirements and can be as accurate as other more expensive methods (Rush, 2001).

A long list of limitations is known about this technique (Rush, 2001), which makes it less reliable to depend on if an accurate estimate is required. Evans (2005) classifies it under the heading "*black-box*" to reflect how inscrutable it is.

This is because it is highly prone to subjectivity, thus considered risky and at risk of error. Three experts with the same starting information may provide different cost estimates, since the use of expert judgement is an unstructured process and not always consistent. Experts are highly prone to bias; their personal experience, political aims, resources, time pressure, memory recall and reasoning are known only to the owner of the estimate.

Estimates depend on the level of experience, are black box in nature and if the experts leave the company their knowledge goes with them. Estimate reuse and modification are difficult and will make effective negotiations with customers very difficult. It is also difficult to provide an audit trail and to quantify and validate the estimates.

Boehm (2000) mentions in his research that the Delphi method can reduce these limitations, due to the nature of this technique, which is considered a sub-set of the traditional expert judgement method or a very useful addition.

2.6.1.4 Delphi Method

The technique was first proposed by Helmer (1967) as a technique for predicting future events; it was developed later into an estimation tool to generate reasonable initial values (Boehm, 2000). The technique is usually carried out in two stages and the following steps (Wu, 1997):

1. Each expert is presented with a specification and an estimation form
2. A group meeting is called in which experts discuss estimation issues
3. Experts fill out forms anonymously
4. A summary of the estimates on an iteration form is prepared and distributed by the meeting chair/leader
5. A group meeting focuses on discussion by the experts of points where their estimates varied widely
6. The experts fill out the forms, either anonymously again or openly (Boehm, 1984), and the last two steps are iterated for as many rounds as appropriate

This iteration method ensures in the end that the result of this procedure has minimum error and is validated.

With the Delphi method, experts can factor in differences between past project experience and the requirements of the proposed project and they can also factor in the project's impacts caused by the modern technologies, architecture,

applications and languages involved in the future as well as exceptional personnel characteristics and interactions, etc.

Some issues inherited from CBR still exist. Typically, this method cannot be quantified; it is hard to document the factors invoked by the experts singly or in their group. Experts can still be biased, optimistic or pessimistic, even though their idiosyncrasies are muted by group consensus.

The expert judgment method always complements other cost estimating methods, such as algorithmic methods (Boehm, 2000; Wu, 1997).

2.6.1.5 Analogy

Employing this technique means "*comparing the proposed project to previously completed similar projects, where the project development information is known*" (Wu, 1997); this is done when there is a lack of information about the new project (Wu, 1997; Ling, 2005). The historical data from the previous project is "extrapolated" to provide a better view in estimating the new project. Wu (1997) considers it a straightforward technique. To achieve an estimate, he advises the following three steps:

1. Characterizing the proposed project.
2. Selecting features from a previous project according to their similarity in characterization.
3. Deriving estimates.

Analogy is very helpful when crucial estimation data is missing (Ling, 2005); analogical estimates are based on actual data with the benefit of being able to use estimators' past experience and knowledge. Most importantly, the

differences between projects can be identified and the impact on future cost can then be calculated. It is, however, similarity dependant; if there is little similarity between projects, it cannot be used.

It is vital to select analogies carefully: too many may dilute the effect and too few will lead to “maverick” projects going forward (Wu, 1997).

2.6.1.6 Cost Estimation with Analytic Hierarchy Process (AHP)

This is a sub-technique from the analogical estimation techniques. It is considered when there are not enough data to produce a detailed cost model (Ling, 2005). It is a systematic technique, mainly employed to aid in decision making. It is based on “experience, intuition and heuristics, the structure of a well-defined methodology derived from sound mathematical principles” (Bhushan and Rai, 2004).

Ling (2005) and Bhushan and Rai (2004) list the six stages needed for a successful estimation using AHP:

1. Breakdown the project into a hierarchy consisting of:
    a. Goal
    b. Sub-criteria
    c. Criteria
    d. Alternatives
2. Data collection from experts corresponding to the hierarchy structure.
3. Organisation of pairwise comparisons in a square matrix.
4. Addition of weights added to the matrix, to show the relative importance of each criterion.
5. Evaluating the consistency of the matrix order *n*.

6. Multiplying each alternative's rating by the weights of the sub-criteria and aggregating them to reveal the local ratings with respect to each criterion.

It is very effective when estimating effort when minimal quantitative data are available (Shepperd and Cartwright, 2001). Nevertheless, since experts are involved, this technique is prone to cognitive bias, optimism, rosy retrospection, underestimation, the subadditivity effect, and bias from memory or lack of experience (Shepperd and Cartwright, 2001; Ling 2005).

2.6.1.7 Regression Analysis

This uses historical cost data to generate a linear relation between old and products costs (Niazi, et al. 2006). It uses statistical approaches and mathematical logic. Normally it is simpler than other techniques, but is only suitable for linear issues (Niazi, et al. 2006).

2.6.1.8 Neural Networks Based Cost Estimation

Neural Networks mimic the human brain, where they can be trained by a set of cost data, which in the case of cost modelling are historic. By training, the computer learns the effect of the product parameters related to cost. This teaches the computer which parameters will influence the final cost (Roy, 2003).

If the neural network was trained with enough historical data, accurate results are obtained (Evans, 2005) and can reveal the relationships between hidden data (Roy, 2003). A neural network can answer questions similar to those of the training data, even if it is handling these questions for the first time (Bode, 2000); moreover, it can easily be re-trained (Bode, 2000). It excels in handling uncertain or non-linear problems (Niazi, et al. 2006; Bode, 2000) and large big numbers of

processing nodes make it more robust. It can also handle fault tolerances (Bode, 2000).

The problem in estimating costs with neural networks is that it is completely data dependent (Niazi, et al. 2006); the user must have enough quality historical data to train the neural networks; hence, on a completely new project it does not function (Roy, 2003) and needs many validated case studies similar to the product, which are not always available, to train the neural networks before they become functional, (Evans, 2005). It is complicated to set up and it functions like a statistical “black-box” (Evans, 2005). Thus, it is also expensive to set up (Niazi, et al. 2006).

2.6.1.9 Parametric Estimating

Parametric models are used as a rule to quantify the unit cost of a product. They do so by employing statistical methodologies and by expressing a product’s cost as a function of its life cycle parameters. These parameters are known as “Cost Drivers” (Niazi, et al. 2006).

Cost Drivers are defined in Nasa’s Parametric Cost Estimating Handbook (1996) as “*the controllable system design or planning characteristics and have a predominant effect on system cost*”. These cost models use a few drivers which have the greatest impact on the product. The drivers are used to establish Cost Estimating Relationships (CERs), which are the equations or algorithms that link the cost drivers to the outputs (Shermon, 2009).

Parametric cost estimation is very effective if the cost drivers are easy to define (Niazi, et al. 2006). It clarifies the influence of different parameters on cost

(Evans, 2005) and is repeatable and objective (Evans, 2005). It can produce excellent predictions if the procedures are followed, the assumptions clearly identified and carefully documented and the data are meaningful and accurate (Roy, 2003).

Using parametric estimation is very sensitive to identified cost drivers; if the cost drivers are not identified, it becomes useless (Niazi, et al. 2006), It too functions like a statistical “black-box” (Evans, 2005) and parameters not included by choice may become important in the future (Evans, 2005).

2.6.1.10 Feature-Based Costing

This is described by Wierda (1991) as “*the integration of CAD/CAM with cost information* [related to the product’s features] *for cost estimation early in the design process*”. It may also be defined as the “identification of cost related features, then the determination of the associated costs” (Niazi, et al. 2006).

To build this cost model only two main steps are needed: identifying the product’s features and determining the associated costs of these features (Roy, 2003; Niazi, et al. 2006). Yet Brimson (1998) sees the procedures to be followed in greater detail. He takes seven detailed steps (Brimson, 1998):

Step 1: Determine the product features

Step 2: Determine the activity routing associated with each product feature

Step 3: Determine the cost of each activity

Step 4: Determine the product characteristics that will cause the process to vary

Step 5: Determine how much the product characteristics cause the process to vary

Step 6: Associate features and characteristics with products

Step 7: Adjust the activity cost based on the product's features and characteristics.

Feature Based costing enables CAD/CAM to be integrated with cost information and can be automated (Evans, 2005). It differentiates features according to cost, thus identifying high cost features (Niazi, et al. 2006). It ties costs and design together (Roy, 2003) leading to better understanding of product costs (Brimson, 1998). It is easy to use since "*Less data is needed to calculate the product cost*" (Brimson, 1998). Therefore, determining and studying features can highlight the factors that cause variation which improves to the product itself or its manufacturing processes (Brimson, 1998).

The main issue with feature based costing is that there is no consensus on what a feature is (Evans, 2005). The more complex and smaller the product is the more complicated the model will be (Niazi, et al. 2006); moreover, it requires large resources to implement (Evans, 2005).

2.6.1.11 Activity Based Costing (ABC)

This technique takes over the modelling tasks from the traditional costing techniques. Starting in the early 1980s, ABC rapidly replaced traditional costing, due to the increased complexity of companies' structures and manufactured products and the rapidly increasing number of products (Andrade, 1999).

This technique estimates cost by breaking down the whole procedure of producing a product into its units, activities. Then afterwards each single activity

is costed separately and the total cost becomes the sum of all the costs consumed by each activity (Niazi, et al. 2006; Roy, 2003).

This method is very simple and effective (Niazi, et al. 2006) and can detect the activities that drive cost up, thus open space for further improvements (Niazi, et al. 2006); but it requires lead time in the early design stages (Niazi, et al. 2006) and needs detailed knowledge of the product and its manufacturing process. These details are usually not available in the early stages (Evans, 2005).

2.7 Research Gap Analysis

In section 2.4 existing LTDP cost models were presented. It was noticed that most cost models were initiated as projects by specific companies or government organisations, like the British library or NASA. Which drove other similar same sector members to generate their own cost models, like CMDP and KRDS. All projects try to follow the OAIS model either closely, like LIFE section 2.4.1, or follow it loosely, like NASA's CET section 2.4.2.

Development of cost models were limited to libraries, i.e. LIFE and CMDP, university and research bodies, KRDS and NASA. The four LTDP cost models went directly into developing targeted cost models, without a path or roadmap on how to generate a similar model for similar sector member, which generated the gap for a framework that enables its users to generate a cost model that serve their respective business sector.

This raises the question, if it's going to be different predicting costs of LTDP systems for business sectors who also are still looking for value versus cost.

The existing cost models reached a single-point cost estimate, which represents the future cost as a single value. This means that little interest was given to the impact on cost generated by uncertainties and obsolescence issues that are within an LTDP system functionality. Obsolescence by nature is an uncertainty (Romero Rojo, 2011) and uncertainties by nature have random probability of occurrence. Therefore, an output of a cost model could be more useful for its users if it reflected those probabilities when and if they impact cost.

Since uncertainties and obsolescence issues were not quantified in previous cost models, they didn't also carry out an uncertainty identification study or an obsolescence study that can identify existing and probable issues within LTDP systems.

There is a lack of cost estimation models for long-term digital preservation activities carried out to serve business sectors other than heritage, libraries and scientific data centres. A full business study of sector-critical requirement differences in long-term digital preservation is lacking in literature and the idea of a cost estimation framework has yet to developed for LTDP. This is due to all cost models were developed directly with directions for developing one's own model. Research on the costs of using cloud computing in long-term digital preservation is also missing. Finally, a lack of knowledge is clear in the areas of uncertainties mitigation costs, obsolescence and the impact of obsolescence on costs in long-term digital preservation. The background to the above aggravate the problem:

- Cost estimation models for LTDP activities in business sectors are missing
- Cloud Computing in LTDP is unique and is currently being introduced
- Analysis of business sector LTDP requirements is critically missing

- Impact of incorporating uncertainties and obsolescence factors on LTDP cost has never been researched
- A framework that could guide users into reaching an LTDP cost model has never been attempted

There for this research project the gap in research is:

1. No research studying cost of LTDP implementation of business sector
2. No framework for LTDP cost estimation
3. Uncertainty impact of cost of LTDP has not been surveyed
4. Obsolescence taxonomy has not been realised

2.8 Summary

Digital preservation research has been the centre of attention in scientific circles that focus on keeping information accessible and usable over long periods. Heritage and scientific data centres paved the way for the development of many cost models and LTDP standards that are nowadays commonly used. Most cost models do not analyse LTDP systems using cloud computer technologies. Cost models developed with a focus on capturing the costs of basic LTDP activities mentioned or inspired by the OAIS reference model. Previous LTDP cost models did not look at the impact on cost estimates of uncertainties and obsolescence issues.

Cloud computing technology was imagined in the early 1960s and was developed until in the late 1990s it emerged as grid computing and then with further development cloud computing became a stable technology with limitless opportunities. The flexible nature of cloud computing gave its technologies a

perfect basis to build on to some innovative LTDP solutions. Cloud computing has three main service models: Software, Platform and Infrastructure. It is deployed in four different models; Private, Public, Community and Hybrid. The main cost elements of owning a private cloud are Power and ventilation, Use and Latency Optimisation, Servers and Computing, Networking equipment and Traffic handling and finally Electrical Power consumption.

Two main categories for cost estimation techniques developed over time; these are directly linked to the designed research methodology, qualitative and quantitative. Qualitative techniques are based on comparing products to check the similarities between old and new products, while quantitative techniques are based on analysing the detailed features, design and activities related to a product. Different cost estimation techniques are suitable for different products in different product development phases. All cost estimation techniques benefit from a solid understanding of the product/service, which cost is being estimated, the development stages and the design.

The ENSURE project, a European funded research project, was initiated to provide a total solution for companies wanting digital preservation in finance, healthcare and clinical trials firms. This research is part of a work package in the European project and aims to generate a framework that can provide a cost modelling solution for long-term digital preservation. The cost estimation process should be specific to the targeted business sectors, while allowing for it to be extended to accommodate other sectors if needed.

3 RESEARCH METHODOLOGY

3.1 Introduction

This chapter aims to discuss the methodology of the research and how it was designed. Each element of the methodology is logically defended, followed by a detailed discussion.

3.2 Research Methodologies and Approaches

A successful research project must follow a strict research method and a design to show how the research itself is to be tackled. This design should define how the research should be broken into clearly defined tasks. A successful methodology should mention the Research Purpose, Research Design, Research Strategy and Data Collection Techniques.

3.2.1 Research Purpose

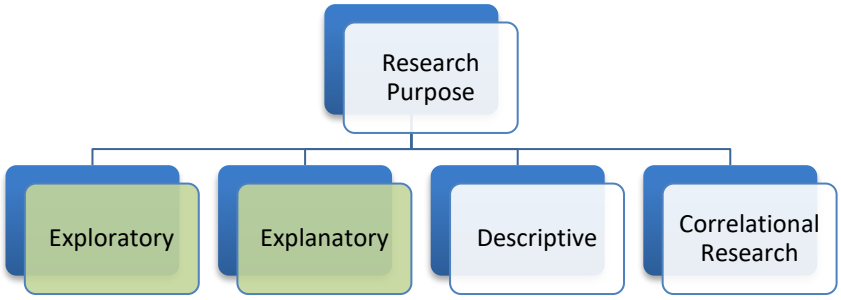


Figure 3-1 Research Purpose (Robson, 2002 and Kumar, 2005)

Robson (2002) classifies research into three kinds according to purpose: Exploratory, Explanatory, and Descriptive; to which Kumar (2005) adds a fourth, Correlational.

- Exploratory aims to discover new insights into what is happening, especially in vague or relatively unknown situations or problems.

- Explanatory kind aims to define or to explain a condition or a relationship in a problem.
- Descriptive aims to describe systematically and accurately the profile of events, people or situations.
- Correlational aims to discover the relationship between two or more aspects of a problem.

3.2.1.1 The Rationale of Utilising Explanatory and Exploratory Approaches

From research project's aim and objectives, it suggests a balanced combination of explanatory and exploratory to be the most suitable methodology approach. Exploratory research leads early stages when knowledge needs to be investigated clearly, not enough knowledge is available, to establish definitions and priorities for the rest of the research(Shields and Rangarajan, 2013) and explanatory research later replaced it in the discrete phases of cost estimation, obsolescence and uncertainty, which enables comparing how cost drivers interact with uncertainties impacts on them and further develop results; i.e. study cause and effect (Brains, et al. 2016).

3.2.2 Research Design

There are two main research approaches, Qualitative or Quantitative (Kumar, 2005).

In quantitative research the data are based, analysed and quantified numerically. The main benefits of quantitative research are that the results are verifiable, controlled by the researcher, replicable and illustrate causal effects. The

downside, however, is that quantitative research is not flexible, it is disconnected from life, disregards experience and interacts minimally with the environment.

In qualitative research, the data are based on observations and discussions. The main benefits of qualitative research are mainly that it responds very well to the world beyond the researcher, factors experiences, contacts participants and studies objects from all sides. Its downside is that it is more exposed to bias than the other approach, its outputs are not directly verifiable, it generates anonymity issues and setting it up takes a long time.

Many authors agree that a design using both approaches ("mixed methods") can support and improve research results (Creswell, 2003); here the benefits of both design approaches can be enjoyed and the disadvantages avoided or at least reduced.

3.2.3 Research Strategy

Some research strategies are suitable for qualitative research and some for quantitative. Others are suitable for a mixed research design, sometimes qualitative and sometimes quantitative.

3.2.3.1 Qualitative Research Strategies

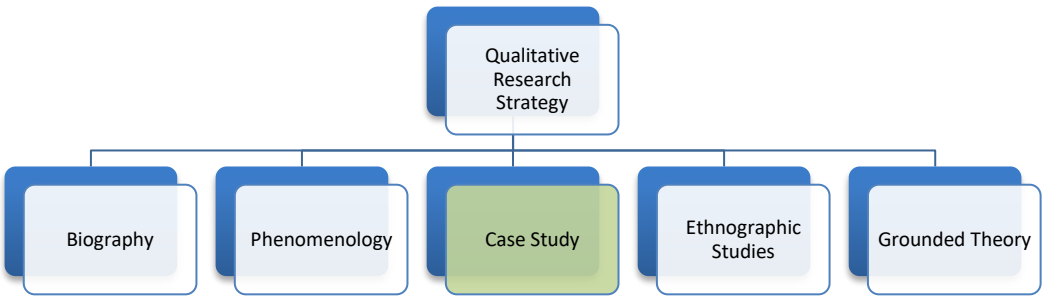


Figure 3-2 Qualitative Research Strategies (Creswell and Poth, 2017)

Biographical research describes the lives of individuals, while the Phenomenological describes people's interactions with a certain phenomenon. The case study strategy, which is very useful for studies that include observations, requires multiple sources of information. Ethnographic Studies concern a collection of people, communities or organisations and their interaction with the world. Grounded Theory is employed when a theory can be generated based on collected data.

The nature of the present research project – that of a multi-organisational project studying LTDP in three different business sectors – made ethnographic studies and case studies suitable strategic choices.

3.2.3.2 Quantitative Research Strategies

Experiments and surveys are the research strategies suitable for quantitative research.

3.2.4 Data Collection Techniques

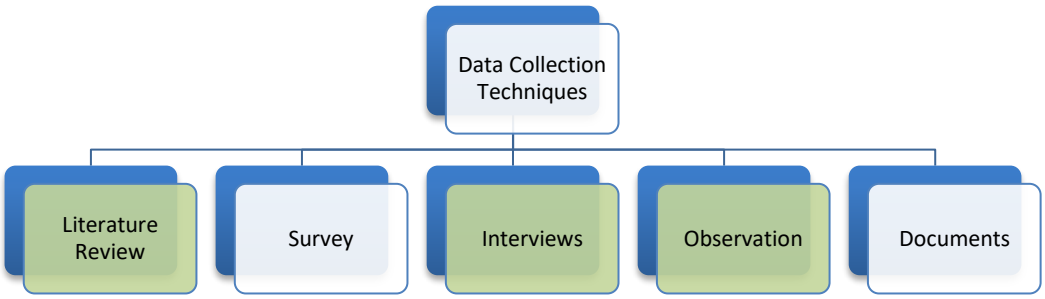


Figure 3-3 Data Collection Techniques (Robson, 2002)

For this research project, the data were collected through a literature review, interviews and observation. The literature review was not only employed to survey the previous relevant work, but rather to find inspiration on how to move forward in subject. Interviews and observation combined were expected to

provide a wide understanding of experts' ideas and opinions and how to carry out the tasks involved.

3.2.5 Methodology Design Summary

In the present research, exploratory and explanatory research designs were combined to strengthen the results as mentioned in 3.2.1.1. The nature of the project led to the following choice of research strategies:

- case studies: the availability of LTDP cost experts enabled the formulation of case studies from the selected business sectors
- the collection of data from:
    - literature review
    - interviews experts
    - observation of systems: one of the major entities that do LTDP to their information and actually some of their experts contributed to the development of the OAIS reference mode, section 2.2.7.1, is available locally in the UK and agreed to let the research in to observe their system first hand.

3.3 Research Methodology Adopted

At this stage, the research was split into three phases which had to be followed to achieve all the required objectives. The phases are a) understanding the context under research and the current state of practice, b) developing the framework and c) validating the framework. A flow diagram of the methodology is shown in Figure 3-4.

3.4 Phase 1 – Understanding the Context and Capturing Current state of Practice

The main goal for this phase is to establish the level of understanding needed to start the research for the project. The starting point for this phase is a combination of four different achievements: first, building the literature review, from reading books and papers from reputable journal and conferences to capture what has been established in this science and what the current state of the art is.

Then the researcher joined mail groups and online communities relevant to the context to ensure a constant stream of news updates on related topics. He also attended introductory courses relevant to the context, for example, introductions to cost engineering, workshops in cloud computing or training in digital preservation. This was to enrich the basic information on the research topic.

The last stage was to familiarise himself with the project's targets and requirements, by reviewing project agreements, work packages and the documents listing requirements. This deepened his understanding of inward and outward scope of the project, and the clear distinction between this project and other research projects; it also helped him to design a suitable research protocol.

Going through these stages ideally results in a deeper understanding of the context, while the practical results in the present case were the ability to design a research protocol and to capture the current state of practice.

The design of the research protocol started with the careful design of a semi-structured questionnaire to capture qualitative and quantitative data, which made it easier to collect the widest scope of information from introductory one-to-one,

phone or web interviews. This questionnaire was piloted by submission to two partners who are already experts in digital preservation activities, to keep the questions relevant and maximise the information extracted by each question.

The design and piloted questionnaire was tried first on introductory visits to two of the expert partners; then used to interview three of the less experienced partners, one from each sector, either face-to-face or by phone or web interviews. This resulted in adjusting the questionnaire according to the interviewee and business sector, thus removing errors and irrelevant questions and maximising the gain from the questionnaire. This was again validated with at least two experienced partners.

Building on this questionnaire to capture the current state of practise, data were collected from four of the partners: one of the expert partners and one from each business sector. Interviews, observation and workshops were employed. The period of observation is still active and will end when sufficient information is gathered from institutes inside and outside the consortium. The main employees interviewed were IT, preservation and/or archival managers, in order to capture the processes and identify the next level of interview candidates. Then the collected data were categorised by business sector and analysed accordingly. This resulted in clear sector differences within the preservation processes. From the three partners three case studies were acquired, to provide a clear understanding of the current state of practice in these business sectors. This completed the stage. These case studies also helped develop the framework.

Finally, all the analysed data and the captured current state were validated again through the expert partners; this was to ensure accurate results in the future and reduce the chance of accumulated deviations.

3.5 Phase 2 – Developing Framework

In this phase two main stages were created to help reach a functional framework. The first of these was to develop an initial framework that could model the cost for digital preservation without going into too much detail, but considered the effect of uncertainties on long-term digital preservation; this was to give a rough idea of what was involved and would pave the way to a more accurate study of uncertainties. To start the initial framework, the work breakdown structure (WBS) and cost breakdown structure (CBS) were developed from the case studies analysed in phase 1. The WBS and CBS were addressed and the full lifecycle of preservation became clearer. This was followed by review of the literature on the cost estimation techniques available, which established which cost estimation techniques would be suitable for the present study.

Finally, an output cost model was constructed. It was generated roughly and quickly by an estimation technique that could give fast and acceptably accurate results, and took the combined knowledge of the full lifecycle, the WBS and CBS, and the careful study of cost estimation. The output test cost model was then validated by two partners with expertise in the area and three partners, one from each sector.

The second and current stage was to develop the framework itself. It started by identifying how many more case studies were needed, with a view to increasing

the design accuracy. Next came a data collection period that involved developing the research protocol with all the industrial partners; some of these companies, according to their industrial sector, were visited and carefully observed, and their activeness in digital preservation was noted. Again, the length of the observation period was determined by the company's experience in LTDP. In this stage the sector's requirements and differences were captured.

Since the effect of uncertainties and obsolescence issues on long-term digital preservation costs is a major question for this research, two actions were required to ensure that these costs were detected, listed and prioritized. Questions on uncertainties and obsolescence had to be an integral part of the questionnaire and then the answers had to be analysed and simulated; hence, an uncertainty identification process was developed along with taxonomy of obsolescence. This was followed by an uncertainty assessment, so as to list all the possible risks and opportunities

The analysed detailed data from all the industrial partners and the uncertainties, opportunity and risk, and obsolescence assessment, were combined to link the cost drivers with the effects of these uncertainties. Finding this link was important in generating a cost model and the final framework.

3.6 Phase 3 – Framework Validation

In this phase the generated cost model and framework will be validated in two stages, through authentic case studies, three of which are an ideal number. Here the framework and its output cost model should be able to estimate the costs with acceptable accuracy. A questionnaire and a workshop were developed then

submitted to expert partners for use in workshops and trials, where they could dictate some examples and, according to their expertise, test the framework and cost model.

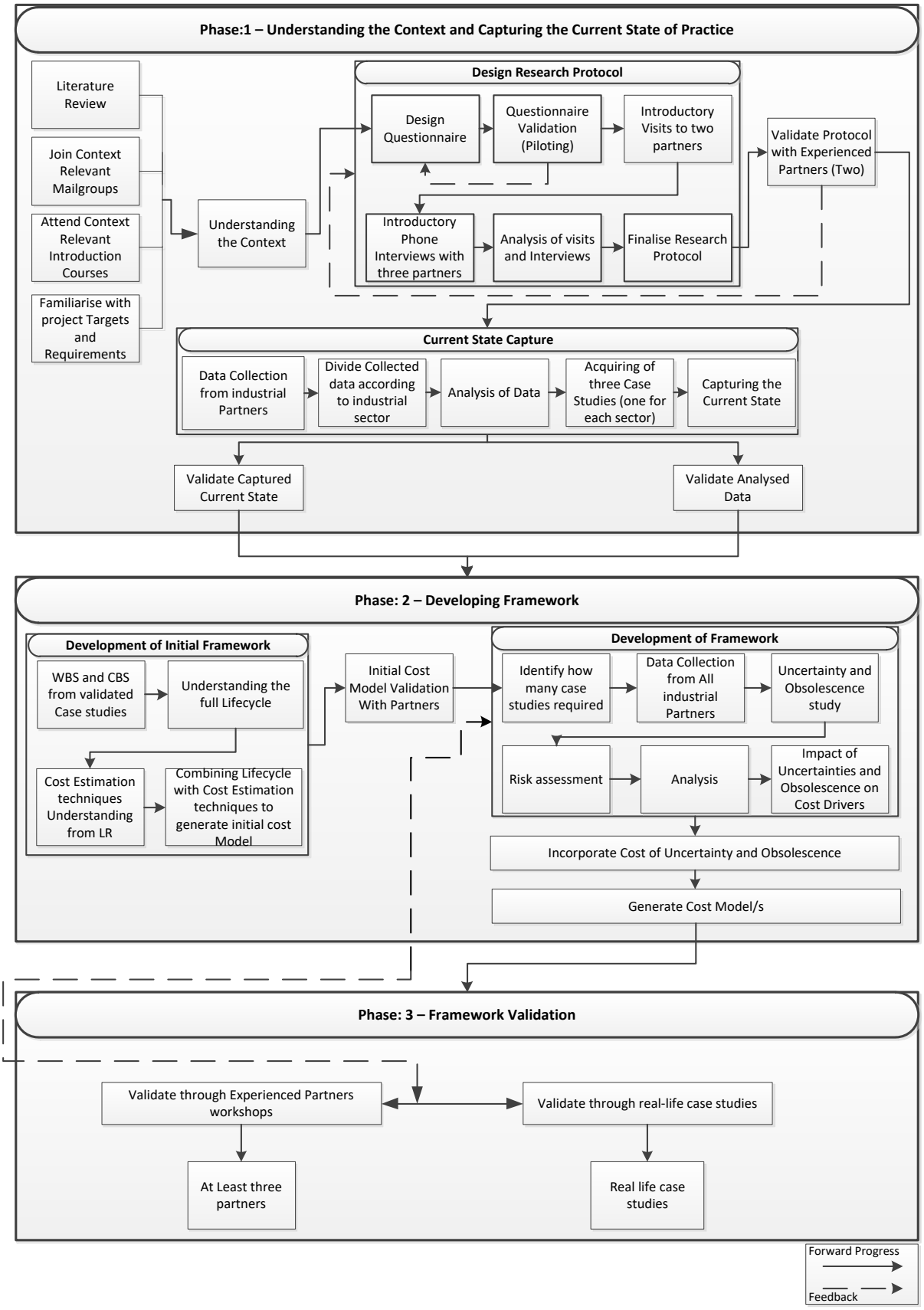


Figure 3-4: Research Methodology

4 LONG-TERM DIGITAL PRESERVATION COST ESTIMATING FRAMEWORK DEVELOPMENT

4.1 Introduction

In Chapters 4 and 5 the development of the framework will be explored in detail. This framework was designed to estimate the whole lifecycle cost of long-term digital preservation (LTDP), considering the impact of uncertainties and obsolescence issues on the estimated cost. The cost estimate was designed around three nominated business sectors: clinical trials, finance and healthcare.

In Chapter 4 the outline of the path to designing the framework, in Chapter 5, and details of enclosed processes are highlighted. Furthermore, the details of the crucial steps in reaching a single point estimate are described.

In the following chapter, Chapter 5, the details of expanding this single point into a three-point estimate are discussed. It requires a full understanding of the way to assess the impact of uncertainties and obsolescence issues on the single point estimate that has been generated. Following this, the construction of the LTDP cost modelling process is described and the framework is constructed.

To reach the main target of this research project, the researcher will start by constructing a cost model for a LTDP system, serving the three targeted business sectors. The three sectors were chosen since all are subjected to the European EU Data Retention Directive (2006/24/EC) and local companies to the UK are subjected to UK Data Retention Regulations 2014 No. 2042; which was a bill in the UK parliament since 2010.

The cost model will predict cost of the LTDP system which nominates cloud computing as the compute and storage solution. This is due to the stability of cloud computing and its future expandability (Foster, et al. 2008; Mell and Grance 2011). This choice was based on the similarities between cloud computing and previous computing and storage solutions, e.g. cluster computing and/or grid computing (Pallis, 2010). This similarity will enable the flexibility of adjusting the cost model to similar technologies and will be ready for current compute technology.

A detailed study of sector's digital preservation requirements and an account of the preservation strategy are presented as a start, highlighting the main cost drivers. The work and cost incurred for a company to produce costing equations based on structured activities and representing the main cost drivers found. These are linked to the work and effort underlying a successful LTDP system.

At a later stage, the uncertainties and obsolescence issues are identified and their impact on cost drivers is linked to each LTDP cost metric. The identified uncertainties and obsolescence factors are linked with their respective cost equations and their effect on cost drivers is defined. A list is compiled of validated assumptions behind some inputs to the cost equations. This is crucial, since cost modelling equations are sensitive to all numerical and logical inputs. This research project's cost model embeds the cost equations and validates the impact of assumptions, uncertainties and obsolescence on cost.

This single point cost estimate combined with the impact of uncertainties and obsolescence on cost generates a three-point estimate (i.e. best-case, worst-

case and most-likely cost) via the use of a Monte Carlo Simulation. The probability of each cost's occurring can be represented graphically on a distribution curve.

4.2 Methodology

The information gathered and analysed to produce the results in this chapter was obtained over three years from rigorous interviews lasting between one and two hours every week with staff from twelve companies. The study were carried out through a series of weekly meetings and interviews with 18 experts from diverse backgrounds who served LTDP systems.

The experts covered all the areas required to fully understand and evaluate LTDP systems serving business sectors under study. A list of the areas of expertise available and the years of experience shared is shown in Table **4-1**.

Table 4-1 Single-Point Cost Model – Experts Details

No.	Company	Area of Expertise of Expert	Experience Years/LTDP
1	IBM	IT Storage and Systems	24
2	IBM	Preservation DataStores	15
3	IBM	Cloud Platforms	32
4	STFC	e-Science	18
5	STFC	Library System	19
6	STFC	Earth Observation Data	10
7	Tessella	Digital Archiving Solutions Development	22
8	Tessella	Digital Preservation Technologies Development	10
9	Custodix	Digital Security Development	11

10	CSISP	Pharmaceutical information	3
11	Maccabi	Healthcare Digital Technology	14
12	Maccabi	Medical Information Systems & Health Records	8
13	JRC	Financial Data R&D	19
14	Phillips	Digital Pathology	7
15	ATOS	IT Systems and Configurations	14
16	Fraunhofer	Healthcare Information Systems	23
17	Luleå University of Technology	Suitability and Quality of Preservation Plans	5
18	University of Porto	Economic Performance	4

Different experts contributed to developing a cost estimating framework and answered many questions targeting specific areas in the design of the framework. Examples of questions asked are:

- What is the volume of the data generated daily?
- What are the activities involved in the preservation phase?
- How long does it take, in man hours, to do (e.g. fixity checks) for a set unit of data?
- What are the activities involved in accessing the preserved data?
- What are the current infrastructure resources available to you for preservation activities?

4.3 Single-Point Cost Model

4.3.1 Introduction to Single-Point Cost Model

The target of this research project is to develop a framework that estimates the whole lifecycle cost of LTDP systems. To be able to reach this target, the researcher has started by developing an LTDP lifecycle then developing a single-point cost model. A single-point cost model produces as single value that represents the estimated cost, without considering any probabilities or skewing to that value. This is a very important initial stage, where the area of research is still vague and more clarity is needed. LTDP businesses requirements, cost drivers and work and cost breakdown structures follow. This will enable any user to develop cost equations, assumptions and rules at this stage and have a functional single-point estimate.

4.3.2 Study the LTDP Lifecycle

It is essential to understand the whole lifecycle behaviour of a digital preservation system, as illustrated in **Figure 4-1**. Additionally, the lifecycle of a digital object in LTDP should also reflect the work breakdown structure realised. To ensure a successful preservation activity, a solid preservation plan must be created. The plan should carefully define the whole journey of the digital content and the activities to undertake if anything threatens the existence or safety of the digital content. All the questions used in this section of interviews are presented in the appendix A.1 to this thesis.

The following steps in the lifecycle are the typical long-term digital preservation activities for planning, submission and ingestion, selecting a type of storage, i.e. public or private cloud, active monitoring of the health of the digital content,

transformation rules, access and retrieval regulations and finally how and when to end the life of a preserved digital object.

Each action in the lifecycle is further described in section 4.3.5, where the lifecycle is expanded into a work breakdown structure (WBS).

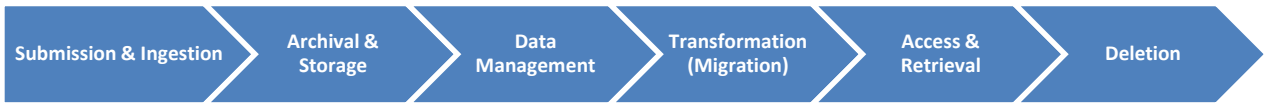


Figure 4-1 LTDP Complete Lifecycle

At this stage, the life of a digital object is defined. Organisations must decide the following:

a. Which file collections deserve the investment of preservation
b. The preservation retention period
c. Access rights
d. Suitable preservation technique/(s) to handle each digital collection
e. Suitable storage and computing technologies suitable
f. The custody requirements of digital objects

Submission and Ingestion are essential steps in the preservation cycle. Selected files should be submitted to the preservation system which by turn will pack them, embed all metadata and preservation information about them and prepare them for storage, preservation and retrieval. The packaging of the digital object follows the requirements of the organisation as reflected in the preservation plan. Indexed and stored for the retention period, digital objects are constantly monitored for any abnormal inhomogeneity in their structure or integrity. When required, a preservation action is taken to protect a collection of digital objects.

The preservation action will be decided according to the nature of the collection and is in the preservation plan. For the present research, migration was thought to deliver the most suitable results for the targeted business sectors. Access rights and means of delivery are considered next, directly before the end of life for a digital object, which is deletion. After extracting the components of the lifecycle of a digital object in a preservation system, the framework user can generate the work and cost breakdown structures.

4.3.3 Identify Sector Differences & Preservation Requirements

A series of interviews with targeted business sectors was conducted to learn their requirements for a digital preservation system. The focus was on learning what they expected rather than what they were currently receiving because this project is targeting new business sectors in digital preservation.

Table 4-2 Sector LTDP Requirements

	Healthcare	**Clinical Trials**	**Financial**
Preservation Duration	Forever, to help with historical big data analysis. Effectively patient age + 25 years	15 years + any Promoter requirements	Client Data = Relation + 5 years. Market Data = 30 years
File Type	Image Video Alphanumeric	Image Video Alphanumeric	Alphanumeric Software
File Format	DICOM* AVI PDF JPEG Doc XML DBA	ECM other PDF GPS* text	PDF Trade DOC station® XLS TXT XPO* (binary)
Access Rate	Very low: once or twice a year	Very low access: maybe none per year and only on inspection. Inspection involves 50 patients' data, varying from 0.5 – 1 GB in each case.	3 yearly audits. 3 cases per audit.
Copy Rights Issues	No copyright issues	Joint ownership with promoter/pharmaceutical company	Software Licence Market Data is not owned
Legal Requirements	For Adults, preserve info for 7 years. For children up to 18 years, preserve info for 25 years.	Data protection act 1599. Contractual agreement with promoter	German BaFin regulations

Questions in appendix A.2 were used to develop the differences in this section. The principal areas of difference, as illustrated in Figure 4-2 and analysed Table 4-2 in were:

a. Preservation Duration (i.e. data retention period)
b. File Types (text, video, audio, images, etc.)
c. File Formats (for each of the preserved file types)
d. Rates of access to preserved files
e. Copy Rights issues (do companies in this sector usually need permission or will any cost be incurred from preserving these digital contents to their owners?)
f. Different legal requirements (any legal obligations on LTDP activities in the sector)

The above sector differences and requirements will affect the total cost of using the LTDP system. For example, legal and copyright issues call for higher security, thus increasing the cost of encryption and decryption processing in ingestion and access activities.

Higher access rates also increase cost, depending on the storage facility used and longer data retention periods will result in significant increases in LTDP costs.

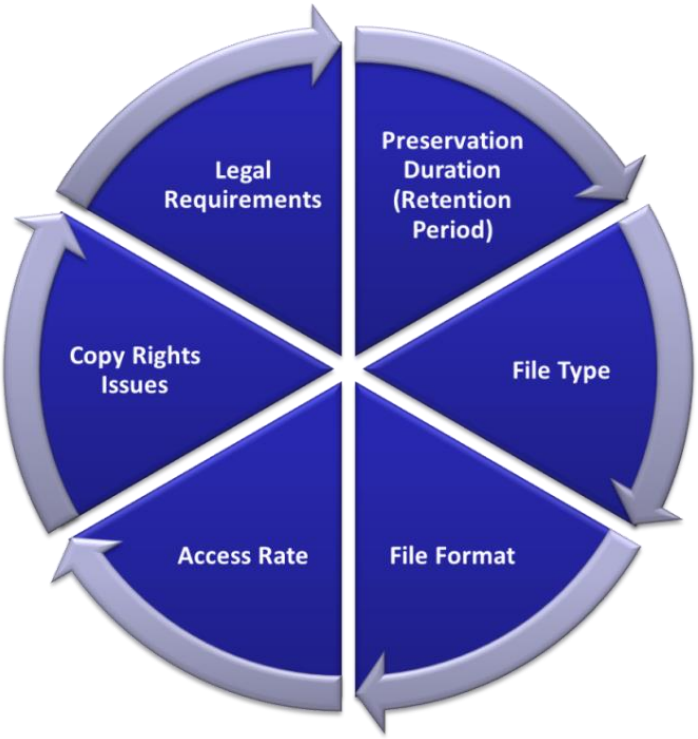


Figure 4-2 Differences in Sectors' LTDP Requirements

As shown in Table 4-2, the answers from business sector experts and from LTDP practitioners, listed in Table 4-1, are collated and compared.

4.3.4 Identifying Key Digital Preservation Cost Drivers

The main cost drivers for the preservation requirements of a business sector should be identified; they depend on sector differences. The main cost drivers that were identified for the LTDP system are the total data volume, data retention period, selection of cloud deployment model (Public or Private) and processing rate of selected IT system.

The total data volume was chosen as the cost driver with the biggest and most direct impact on total cost, since all the staff, computing and storage requirements are purchased or rented simply to accommodate this volume. The next driver in

impact is the retention period; the longer the retention period the higher the cost, due to the longer commitment to maintaining the LTDP system and staff.

The cloud deployment model has a major impact on the initial costs, which may either be too high with a private cloud or comparatively low with a public model. The IT system purchased/rented must be chosen to a high degree of accuracy with continuous adaptation to the needs of the LTDP system over time. If the system is too slow it consumes too much time and power to finish the requested job, while if the system is too fast for the job it will remain idle for too long thus consuming power with no actual benefit.

4.3.5 Construct Work and Cost Breakdown Structures

The digital preservation activities in the work breakdown structure (WBS) are broken down into five main activities. Pre-ingest and ingest are triggered by the submission of new digital objects. In these activities, the object is prepared and packaged for preservation. Data management is the activity that is designed to monitor all aspects of the LTDP system; it is always active and it co-ordinates the other activities. Access handles delivery of the files to the requesting user and finally Transformation action handles any requested preservation action in the LTDP system. Figure **4-3** shows a high-level representation of WBS for the LTDP system of this research project.

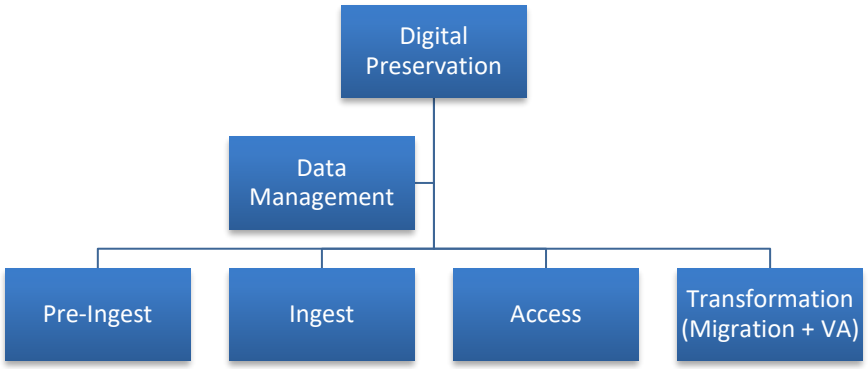


Figure 4-3 LTDP System High Level WBS

The detailed breakdown of each activity is important for understanding the flow of work and effort expended. The WBS and CBS discussed in this section was derived from questions in appendix A.1 and from literature as seen in section 2.2 of this document.

4.3.5.1 Work Breakdown Structures

a. Pre-Ingest

As is clear from its name, this activity occurs before the ingestion of data; it is where the Submission Information Package (SIP) is prepared alongside the generation of metadata (see Figure **4-4**). The complete work breakdown structure can be seen in appendix A.7.

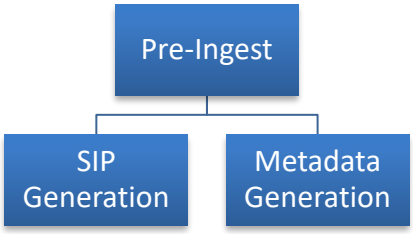


Figure 4-4 Pre-Ingest WBS

Although it is not a part of the OAIS reference model, LTDP systems designers, experts number 2, 4, 5, 7, 8, 11, 12 and 15 from Table 4-1, highly recommend

separating the two functionalities. The separation enables the pre-ingest phase to target negotiating with human factors and helps to ensure that the data collected are suitable and validated.

b. Ingest

Ingest, shown in Figure **4-5**, comes directly after pre-ingest. The first thing it does is to make an initial fixity check, which is used later by data management for the scheduled fixity checks. Then a quality control procedure follows, with a virus check to make sure that the files ingested are virus free.

Quality control also ensures that the ingested files are permissible and do not interfere with any pre-set regulations. Afterwards file format identification and validation are carried out to make sure that the files format extensions match the construction of the files themselves. Then the file properties are extracted and added to the metadata.

If the ingested files contain other embedded options or other embedded files, these are extracted and then ingested as new files but linked through their metadata to their parent file.

The ingest activity sends the files packaged and RDF annotated to the storage facility and writes their locations to the file index table and also, to simplify search and retrieval, sends these to the metadata as an indexer, together with the location of their backups and the location of their metadata. The ingest activity also has a data protection action, which sets the user's authorisations and encrypts all the ingested files, according to their security requirements.

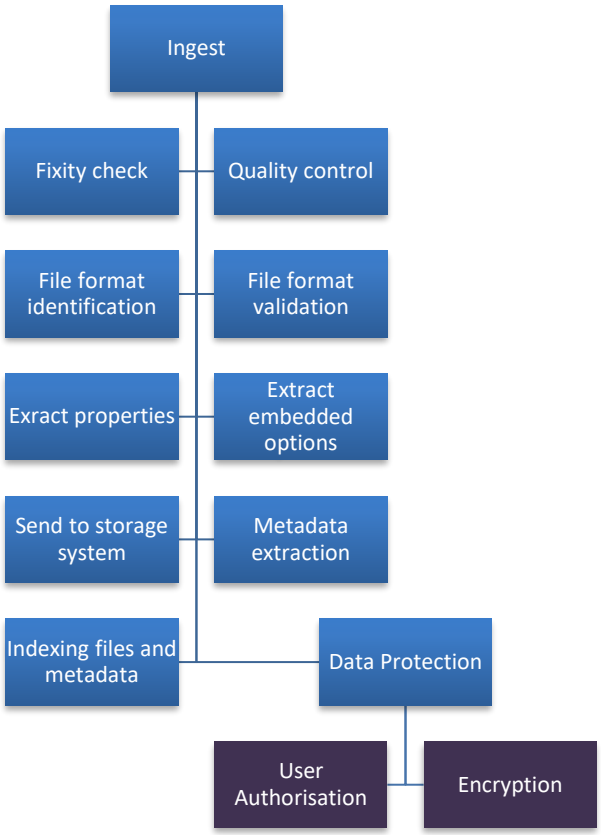


Figure 4-5 Ingest WBS

c. Data Management

This activity is always active, since it monitors data continuously and takes action according to a schedule of tasks, an emergency or a customer's requirements. Fixity checks are usually hooked into a schedule, so after every certain interval a fixity check is performed to confirm that the data bit streams are intact. If not, data management will issue a copy command from the intact backup. Appraisals come from the customer, for the preservation of a file or set of files does not mean that these files must be preserved forever. The customer can make an appraisal to request an Authorised Deletion of an old dataset, or the Edit/Update of the accompanying metadata.

The authenticity and provenance of the digital objects are handled in data management, ensuring the *trustworthiness and history of creation, ownership, accesses and changes* (Factor, 2009). Transfer of custody can also be requested by the user, to move the data sets from one handler to another. Data management also handles accessibility, and gives orders to the access activity (see Figure **4-6**), to allow or forbid access to the files according to a set of accessibility options. The characterisation of a new ingested file, or a file that needs to be migrated to a new format, and the reporting functionalities are also handled by Data management. All other activities in this WBS are sequential; they are triggered by input and switched off by output.

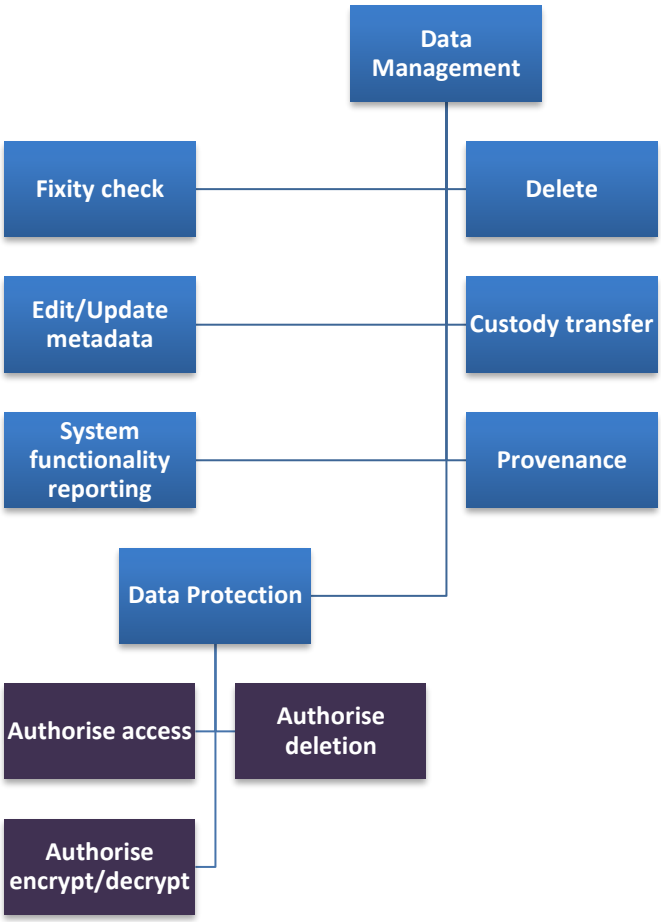


Figure 4-6 WBS of Data Management

d. Access

When the user requests access to a file or a set of data, s/he does it through a search engine that checks through the file indexing and metadata, where the Ingest activity recorded the location of the stored files.

After Data Management gives the order to the Access activity (see Figure **4-7**), the file is retrieved and offered to the user as a download package from the storage facility through the LTDP's system web interface or in the interface via the virtual appliance. Access Activity includes some data protection actions, to authorise all access activities and to anonymise them, decrypt accessed files and carry out security filtering.

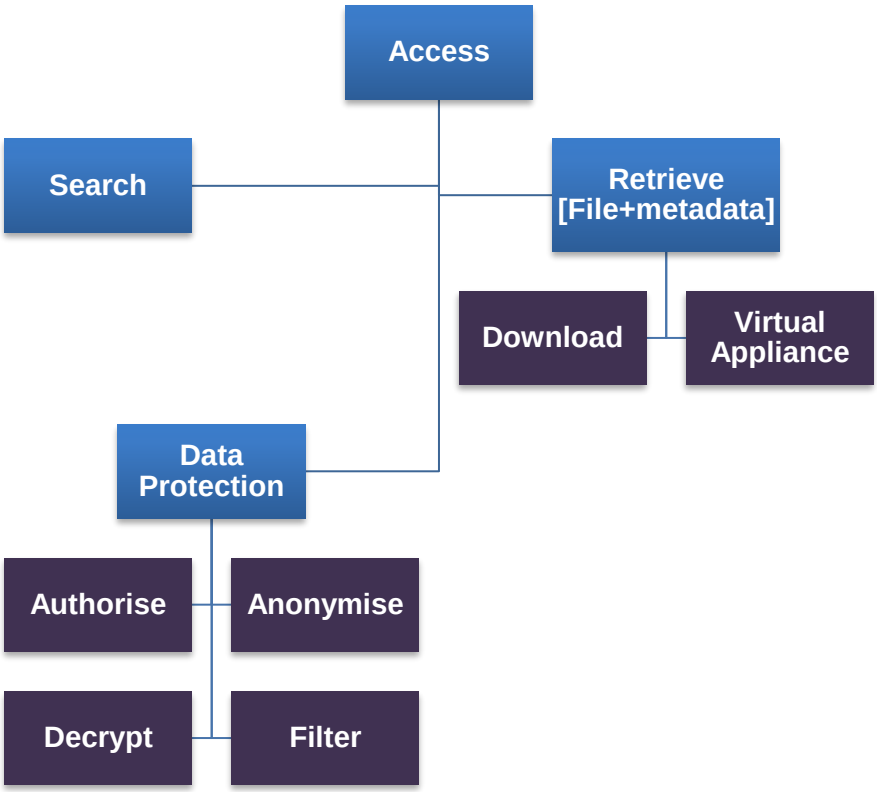


Figure 4-7 WBS of Access

e. Transformation

Transformation Management, Figure **4-8**, is usually idle, unless it receives a command about a specific file or data set. The system activates this activity, to either migrate or initiate a virtual appliance. If the command received is to Migrate then it Retrieves the files to be migrated to the new file format, migrates them, compares them to the original file and then sends the new file format to be Re-ingested again as a new file, while keeping the old files for reference.

However, if a Virtual Appliance is activated; the required Appliance becomes active on the computing facility and the files are provisioned to and from the appliance. In both cases data are always protected via authorisation checks, and then, if these clear the request, the encrypted files are decrypted.

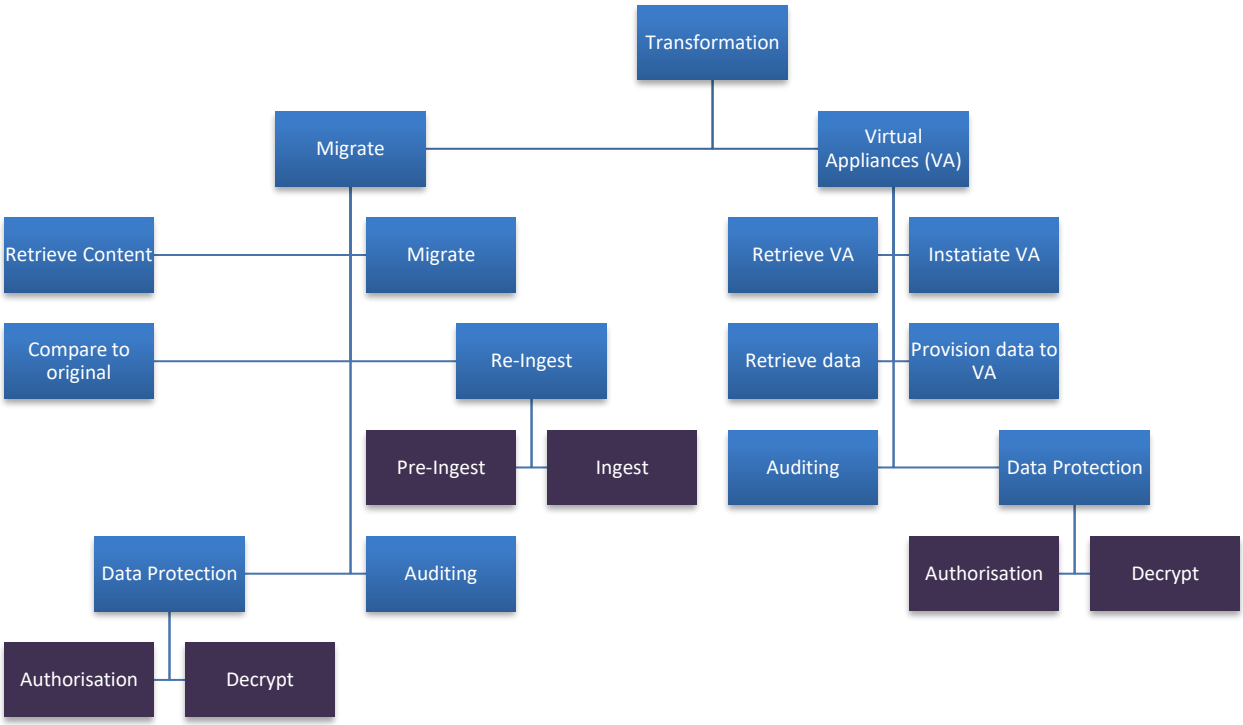


Figure 4-8 WBS of Transformation

4.3.5.2 Cost Breakdown Structures

After generating the WBS for all preservation activities, the Cost Breakdown Structure (CBS) can be derived from the WBS, as illustrated in Figures 4-9 and 4-10. This shows the detailed cost breakdown for the Compute and Storage cloud deployment models. Detailed cost breakdown for a Private Cloud-based LTDP solution is shown in Figure **4-9**, while detailed cost breakdown of a Public Cloud based LTDP solution is shown in Figure **4-10**; where it is assumed that service providers for public cloud are Amazon® (AWS® Services) and Rackspace®.

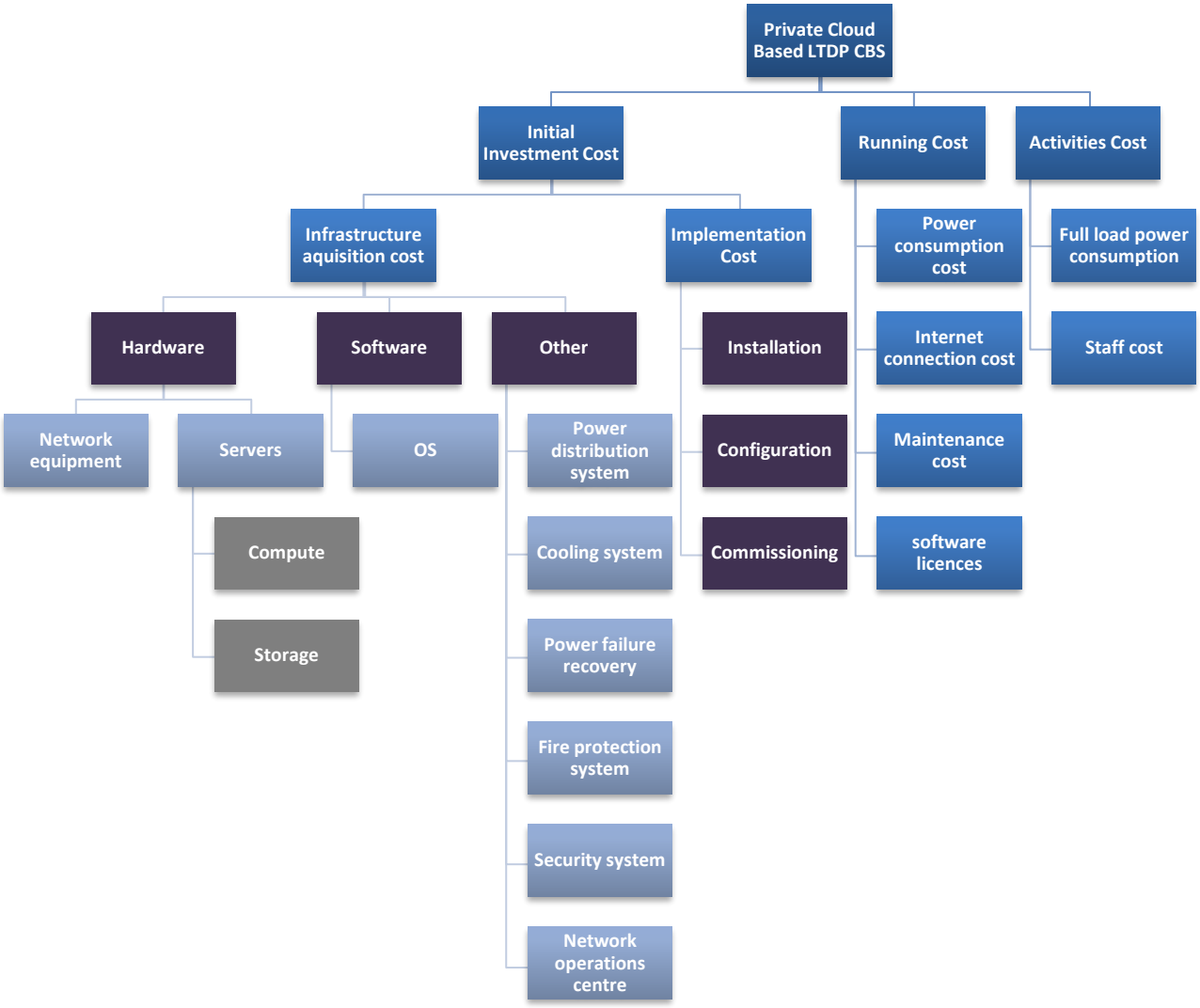


Figure 4-9 Detailed CBS of a Private Cloud Based LTDP Solution

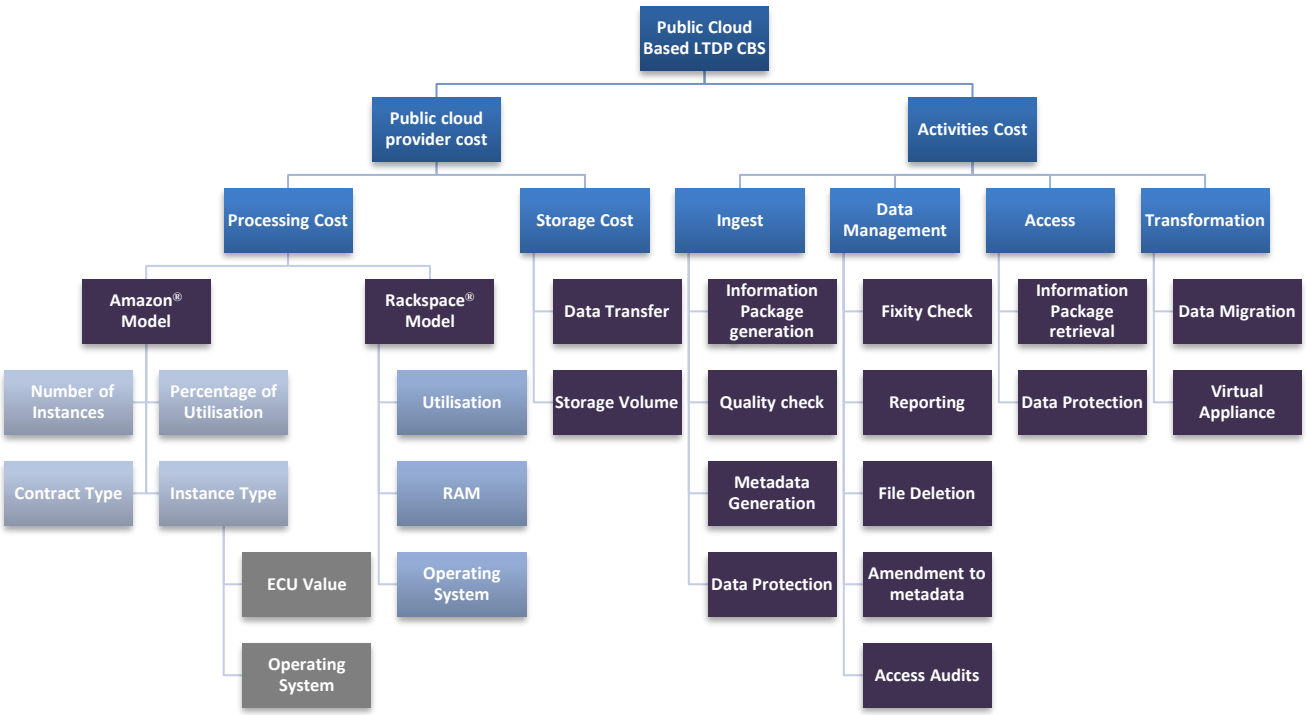


Figure 4-10 Detailed CBS of a Public Cloud Based LTDP Solution

4.3.6 Generate Cost Equations and Rules

All the CBS elements from the previous section was broken down into equations, which were designed to reflect the behaviour of the LTDP system. The main cost equations, developed for the cost engine, are split into two main parts. These are the private cloud deployment model and public cloud deployment model. Both models followed their own cost breakdown structure as illustrated in Figure 4-9 and Figure 4-10. Complete cost breakdown structure is shown in appendix A.8.

Below are details of the main cost equations used to calculate the total cost of using the preservation system on both cloud deployment models. Any assumptions made are discussed and explained following the cost equations.

4.3.6.1 Private Cloud Based Cost Model

The private cloud based LTDP system cost model (*PrivCM*) is designed based on information gathered from the literature, section 2.2, where LTDP techniques and OAIS reference model where discussed and from the CBS in **Figure 4-9**. It is composed of three main elements, as shown in equation **4-1**.

Initial investment cost (*iiC*) represents the initial investment that a firm needs to commit to obtain the LTDP system and start it up. Running cost (*RunC*) represents the costs generated when the LTDP system is running idle with no LTDP activities, while keeping the system maintained.

Finally, the activities cost (*ActC*) is the cost generated by a trigger when one of the preservation activities is requested/started by the LTDP system or user.

$$PrivCM = iiC + RunC + ActC$$

4-1

I. Initial Investment Cost

$$iiC = infaqC + imC$$

4-2

where *iiC* is initial investment cost, *infaqC* is the infrastructure acquisition cost and *imC* is the implementation cost, which represents all incurred costs of implementing, installing and commissioning the new system.

- **Infrastructure Acquisition Cost**

$$infaqC = HwC + SwC + OtC$$

4-3

where *HwC* is the Hardware cost, *SwC* is the Software cost and *OtC* is other costs. Other costs include all the peripheral systems that will serve storage and the compute (IT) system.

- **Hardware Costs**

$$HwC = NetEquipC + ServC + StrMC + TermC$$

4-4

where *NetEquipC* is the Network Equipment Cost, *ServC* is the servers cost, Compute and Storage, *StrMC* is the Storage Media cost and *TermC* is the cost of terminal computers (each terminal was set to € 750).

$$NetEquipC = RoutC + SwtcC + ApC$$

4-5

where *RoutC* is the Routers cost, *SwtcC* is the switches cost and *ApC* is the Access point cost.

$$ServC = CSC + SSC$$

4-6

where *CSC* is the compute servers cost and *SSC* is the storage servers cost (SSC is the number of storage servers multiplied by cost).

$$CSC = FFMC + PMC + RAMC$$

4-7

where *FFMC* is the form-factor model cost, *PMC* is the Processor model cost and *RAMC* is the random-access memory size cost.

$$StrMC = TapRobC + (TC + HDDC)$$

4-8

where *TapRobC* is the tape robot cost, *TC* is the tapes cost and *HDDC* is the hard disk drive cost.

- **Software Costs**

$$SwC = OSC + SPC$$

4-9

where *OSC* is the Operating systems cost and *SPC* is the Software purchase cost.

Software purchase includes both purchased and developed software costs (*DevSwC*).

$$DevSwC = DevTime \times \#\, Developers$$

4-10

$$OSC = ClCOSC + ServOSC$$

4-11

where *CLCOSC* is cloud computing operating system cost and *ServOSC* is Servers Operating system costs.

- **Other Costs**

$$OtC = PDSC + CoolC + PFRC + FPC + SecuC + NOCC$$

4-12

where *PDSC* is the power distribution system cost, *CoolC* is the cooling system cost, *PFRC* is the cost of the power failure recovery system, *FPC* is the fire protection system cost, *SecuC* is the security system cost and *NOCC* is the network operation centre costs.

$$PDSC = PDuC + TransC + PCMuC + CabC + CabTC$$

4-13

The power distribution system costs, as shown in Figure **4-11**, consist of the power distribution unit costs, *PDuC*, the transformers' cost, *TransC*, the power consumption monitoring equipment cost, *PCMuC*, cables cost, *CabC*, and the cost of the cable trays that hold all the cables, *CabTC*.

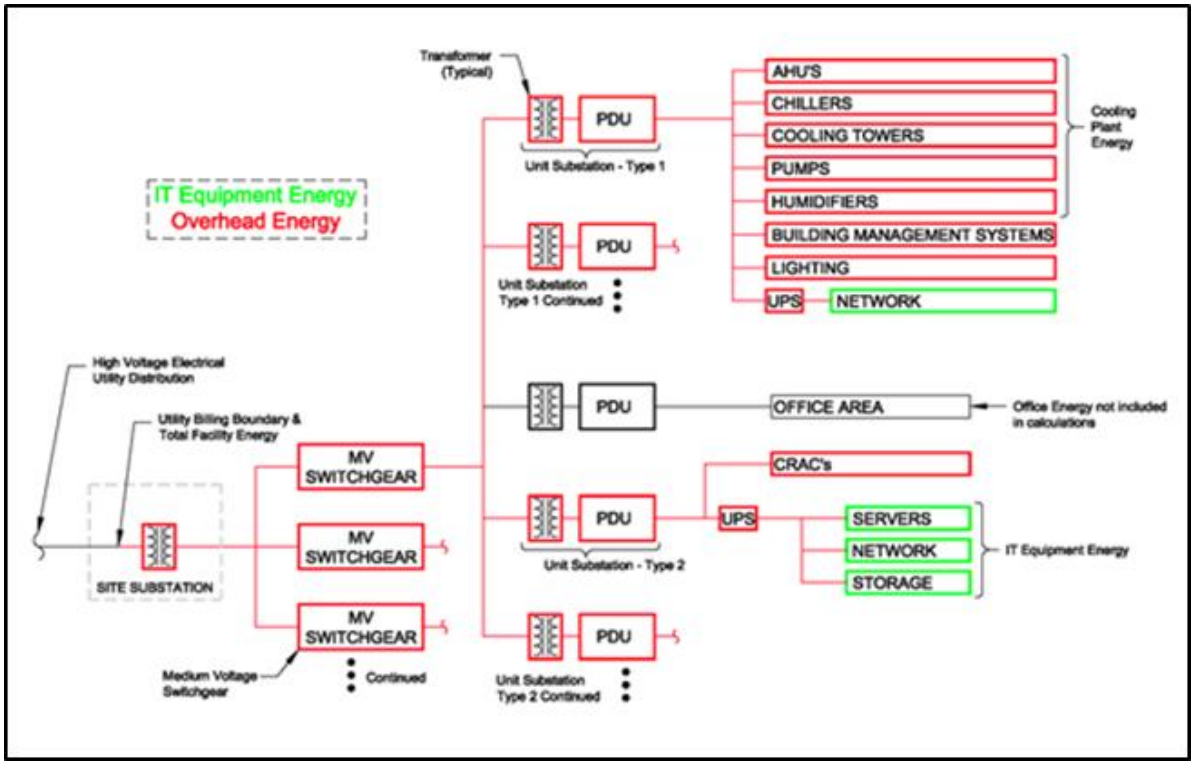


Figure 4-11 Google Data Centre's Power Distribution Schematic (Google, 2011)

$$CoolC = AHuC + ChillC + CoolTC + PumpC + HumidC + CRAC$$

4-14

The cooling system costs, as shown in Figure **4-12**, consist of air handling units' cost, *AHuC*, the chillers cost, *ChillC*, the cooling towers cost, *CoolTC*, together with the cost of Pumps, *PumpC*, humidifiers, *HumidC*, and computer room air conditioning units, *CRAC*.

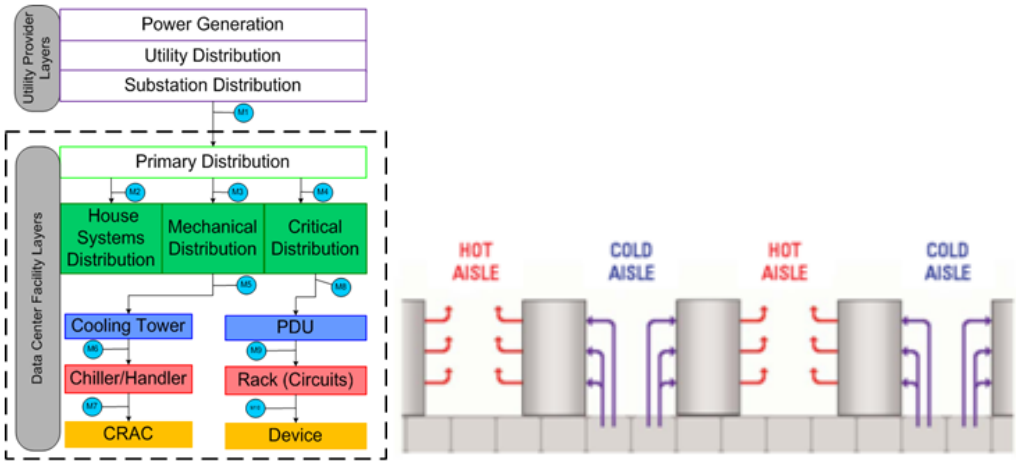


Figure 4-12 Schematic of Power and Cooling Systems in a Data Centre (Microsoft 2008)

$$PFRC = UPSC + EDGC + ATSPC$$

4-15

The Power failure recovery system costs consist of an uninterrupted power supply cost, *UPSC*, emergency diesel generator costs, *EDGC*, and the automatic transfer switch panel cost, *ATSPC*.

$$FPC = DetC + FPSC$$

4-16

The fire protection system costs consist of costs corresponding to the number of detectors used, *DetC*, and the fire suppression system, *FPSC*.

$$DetC = Smoke + Heat$$

4-17

The fire system Detectors consist of smoke and heat detectors.

$$FPSC = GasCC + PipeC + MonSC + CabsFC$$

4-18

The fire suppression system consists of gas cylinders, *GasCC*, pipes for the gas delivery, PipeC, a monitoring system, *MonSC*, and fire cables, *CabFC*.

$$SecuC = FwC + CCTV + ASC + ACCS$$

4-19

The security system is composed of a firewall (software), *FwC*, CCTV monitoring system, *CCTV*, alarm system, *ASC*, and access control system, *ACCS*.

$$NOCC = NOCMonC + CommEqC$$

4-20

The network operations centre is where staff monitor the performance of the data centre and consists of the monitoring system, *NOCMonC*, and communication equipment, *CommEqC*.

- **Implementation Costs**

$$imC = InstalC + ConfigC + ComC$$

4-21

where *InstalC* is installation costs, *ConfigC* is configuration costs and *ComC* is commissioning costs. All are calculated by multiplying the hourly cost of staff by the number of staff required by the total number of hours.

II. Running Costs

$$RunC = PCC + MainC + AnSwC + RunNOC$$

4-22

Where *PCC* is the power consumption cost for the whole system except the full load of the IT system. *MainC* is the maintenance cost and *AnSwC* is the annual software licences cost.

$$\mathrm{PCC} = (IdleConC + CoolPwC) \times kWhC$$

4-23

Power consumptions costs are the idle power consumption of the IT system, *IdleConC*, added to the cooling system power consumption, *CoolPwC*, and then multiplied by the cost of one kilowatt hour, *kWhC*.

$$MainC = Prevent + Correct$$

4-24

The costs of maintenance are divided between two main maintenance strategies, preventative and corrective maintenance, *Prevent* and *Correct* respectively.

$$Prevent = (CSCRefC \times RefCompR) + (SSCRefC \times RefStorR)$$

4-25

Preventative maintenance costs are the compute servers refreshing costs, *CSCRefC*, multiplied by the refresh rate of the compute servers, *RefCompR*,

added to the storage servers refreshing costs, *SSCRefC*, multiplied by the refresh rate of the storage servers, *RefStorR*.

$$Correct = (HDDC \times HDFailR) + (TC \times TFailR)$$

4-26

The corrective maintenance costs are the expected failure rate of hard disc drives, *HDFailR*, and tapes, *TFailR*. The expected failure rate is then multiplied by the corresponding cost of the storage medium.

$$RunNOC = NOCStf \times (SOh \times 2080) \times MonTime$$

4-27

Running the network operations centre incurs costs in staff time. This is broken down into the number of staff required, NOCStf, multiplied by their hourly salary including overheads, SOh, multiplied by the number of hours per year, 2080. Finally, the previous total is multiplied by the total time dedicated to monitoring, MonTime.

III. Activities Cost

$$ActC = \left(\frac{DV}{ITspR}\right) \times ((ITflC \times kWhC) + (Staffact \times SOh))$$

4-28

where *DV* is the data volume that the activity will be requested to handle. *ITspR* is the IT system processing rate and the speed of the IT system at performing the activity per one unit of data. *ITflC* is the power consumption of a fully loaded IT system. *kWhC* is the cost of one kilowatt hour. *Staffact* is the number of staff required for the activity and *SOh* is the salary and overheads per hour per

member of staff. The assumptions, behind this cost model's equations, are discussed below in Chapter 6.

4.3.6.2 The Public Cloud Based Cost Model

The public cloud based cost model constructs the cost using two main elements, activities and the provider's cost structure. Whenever an activity is triggered, it generates costs from using the cloud computer, while uploading any gigabyte (GB) into the preservation system storage generates costs from using the provider's storage facilities. Unlike the private cloud solution, where the client commits initial funds and then maintains the system, here the client pays for what is used and is billed monthly.

The nominated cloud providers are Amazon's AWS® and Rackspace®. The selection was based on their reliable services and products, which mainly target professional users. They also heavily invest in improving their current technologies while constantly providing new solutions. Calculating the total cost of LTDP by public cloud depends on understanding how the provider builds up its billing mechanism. The following equations show how to calculate compute and storage costs for each of the nominated cloud providers.

I. Cloud Providers Costs

Both providers have similar storage costing mechanisms and follow equation **4-29**. Cloud storage cost is mainly generated by the data volume of the preserved material, *DV*, unit cost of storage for the selected cloud provider, *CStrC*, and the duration of the preservation, *Dur*. To the previous basic cost of cloud storage, the

costs of traffic generated by the number of specific requests by the user, *NReq*, multiplied by each request price, *ReqC*, are added.

$$CStrC = (DV \times uStrC \times Dur) + (ReqC \times NReq)$$

4-29

There are four types of request to a cloud service: data transfer in and out, *DTinC* and *DToutC* respectively, refer to the data moving in or out of the server. Put/Copy/Post/List, *pcplRC*, requests are related to publishing the stored objects. Get/other, *goRC*, requests are related to pulling the object from the server. The cost of requests is calculated by adding all the requests costs together, as shown in equation **4-30**.

$$ReqC = DTinC + DToutC + pcplRC + goRC$$

4-30

The data transfer in cost, *DTinC*, is the cost of adding files to the cloud storage and is calculated as the data volume, *DV*, multiplied by the data volume in price, *DTinP*. Most providers keep this cost at zero simply to encourage users to add files to their library of files.

$$DTinC = DV \times DTinP$$

4-31

The data transfer out cost, *DToutC*, is the cost of extracting files from the cloud storage. It is exactly similar to data transfer in, but the provider asks a price for transferring data from its facilities, *DToutP*.

$$\boldsymbol{DToutC = DV \times DToutP}$$

4-32

Put/Copy/Post/List cost is the number of these requests, *NpcplR*, multiplied by their cost, *pcplRP*, usually priced when requests reach over 10,000 per year, equation **4-33**.

$$pcplRC = NpcplR \times pcplRP$$

4-33

Similarly, the GO/other requests cost is also calculated as the number of requests, *NgoR*, multiplied by the request cost, *goRP*. These requests incur a price when they reach over 100,000 requests per month, equation **4-34**.

$$goRC = NgoR \times goRP$$

4-34

- **Amazon Compute Processing Cost**

Amazon® AWS® has a pricing scheme to the compute servers that is broken down into several elements. Each initiation of an action on a machine is called an instance; instances can be initiated simultaneously. Each of the machines has a specific processor, RAM, internal storage which is different from the main storage where preserved objects are archived and a different operating system; Windows® or Linux/UNIX®. AWS® ranks cloud computers in ECU units.

Elastic Compute Cloud (EC2) Compute Unit (ECU) is the indicator of how powerful a cloud computer is. AWS® defines ECUs as follows: "*The amount of CPU that is allocated to a particular instance is expressed in terms of these EC2*

Compute Units" (Amazon, 2014). Those machines are either used on a usage percentage basis or on a 100% utilisation contract.

To find the processing cost, as shown in equation **4-35**, over an AWS® machine, *AmznPC*, the number of instances, *NInsta*, is multiplied by the utilisation percentage, *Utlcnt*, and the cost of the Instance contract type, *InstaC*.

$$AmznPC = NInsta \times Utlcnt \times InstaC$$

4-35

- **Rackspace® Compute Processing Cost**

Rackspace® has designed its pricing, *RksPC*, around the hourly rate of a machine with specific OS and RAM size, *OSuhrC* and *RAMuhrC* respectively. Two choices of OS are available, Windows® and Linux®; with three RAM sizes for Windows® and four RAM sizes for Linux®. This gives the client seven options to choose from in relation to the number of servers needed, *NServ*, and the number of hours of use per month, *Nhrs*. Finally, the expected traffic cost is added, *SrvTrfc*, which is a fixed value per GB. Equation **4-36** shows how to calculate the cost of using Rackspace® compute cloud service.

$$RksPC = \left((OSuhrC + RAMuhrC) \times NServ \times Nhrs\right) + SrvTrfC$$

4-36

Each LTDP activity has a different impact on cost. The cost equations for each activity are designed to express what WBS previously discussed. The core of each activity cost is the processing rate, namely, how fast the computer will process the preserved objects through the active LTDP activity.

II. Activities Cost

- **Ingestion Cost**

Ingest cost, *IngC*, as shown in equation 4-37, is the sum of the cost of generating the submission information package, *SIPgC*, the quality check cost, *QCC*, the description data (metadata) generation cost, *metagC*, and the Data protection for Ingest cost, *DPrtingC*.

$$IngC = SIPgC + QCC + metagC + DPrtingC$$

4-37

The Submission information package generation cost is generated as the processing rate, *ingPR*, of the submitted data volume, *DV*, multiplied by the processing cost related to the nominated cloud provider (*), **PC*.

$$SIPgC = \frac{DV}{ingPR} \times * \text{ PC}$$

4-38

The quality check cost is calculated as the sum of metadata extraction cost (*metexC*), metadata validation cost (*metVlidC*), file format identification cost (*filidC*), single fixity check cost (*SfixC*) and file properties extraction cost (*filprpC*)

$$QCC = metexC + metVlidC + filidC + SfixC + filprpC$$

4-39

All the cost elements of QCC have a similar equation, $Cost\ Element = Data\ Volume \times {}^{Provider\ Processing\ Cost}/_{Cost\ Element's\ Processing\ Rate}$. In

Equations **4-40**, **4-41**, **4-42**, **4-43** and **4-44**, all cost elements equations show different specific cost element processing rates.

$$\boldsymbol{metexC} = \frac{\boldsymbol{DV}}{\boldsymbol{metexPR}} \times * \boldsymbol{PC} \qquad \textbf{4-40}$$

$$metVlidC = \frac{DV}{metVlidPR} \times * PC \qquad \textbf{4-41}$$

$$\boldsymbol{filidC} = \frac{\boldsymbol{DV}}{\boldsymbol{filidPR}} \times * \boldsymbol{PC} \qquad \textbf{4-42}$$

$$\boldsymbol{SfixC} = \frac{\boldsymbol{DV}}{\boldsymbol{SfixPR}} \times * \boldsymbol{PC} \qquad \textbf{4-43}$$

$$\boldsymbol{filprpC} = \frac{\boldsymbol{DV}}{\boldsymbol{filprpPR}} \times * \boldsymbol{PC} \qquad \textbf{4-44}$$

- **Data Management Cost**

The data management cost (*DMC*), equation **4-45**, is the sum of fixity check costs (*FixC*), reporting cost (*Repc*), file deletion cost (*FildelC*) and amendments to metadata cost (*AmmetC*).

$$DMC = FixC + RepC + FildelC + AmmetC \qquad \textbf{4-45}$$

The fixity check cost, equation **4-46**, is the total cost of every fixity check made after the initial one done on ingestion. It is the fixity processing cost multiplied by the annual frequency of checks (*AFreqfix*) and the duration of the preservation period (*Dur*).

$$FixC = (\frac{DV}{FixPR} \times * PC) \times (AFreqfix \times Dur)$$

4-46

The audit reporting cost (*RepC*), equation **4-47**, multiplies the processing cost of generating the reports with the frequency of system reporting (FreqRep).

$$RepC = \frac{RepDV}{RepPR} \times * PC \times FreqRep$$

4-47

The file deletion cost (*FildelC*) is calculated by multiplying the cost of deletion requests (*delRqC*) by the number of deletion requests (*NdelReq*).

$$FildelC = delReqC \times NdelReq$$

4-48

The metadata amendment cost (*AmmetC*), equation **4-49**, is the product of the number of amendment requests (*NAmmet*) and the Get request cost (*goRC*). The result is added to the cost of re-ingesting this data volume of metadata (*metDV*)

$$AmmetC = (NAmmet \times goRC) + ingC_{metDV}$$

4-49

- **Access Cost**

Accessing data is the main reason for carrying out LTDP activities. Access cost (AcC), equation **4-50**, is the total of adding the cost of dissemination information package retrieval (*DIPretC*) to the cost of data protection for access (*DIPrtAC*).

$$AcC = DIPrtC + DPrtAC$$

4-50

The dissemination information package retrieval cost, equation **4-51**, is calculated by adding the cost of processing the information package data volume to the cost of data out transfer (*DToutC*).

$$DIPrtC = \left(\frac{DV}{DIPrtPR} \times * PC\right) + DToutC$$

4-51

The data protection for access cost, equation **4-52**, is the cost for data decryption.

$$DPrtAC = \frac{DV}{DPrtACPR} \times * PC$$

4-52

- **Transformation Cost**

The transformation cost (*TransC*), equation **4-53**, or cost incurred for taking preservation action, to migrate (*DMgrtC*) and/or use a virtual appliance (*VApinC*) on the cloud compute server thus enabling the user to retrieve preserved data.

$$TransC = DMgrtC + VAC$$

4-53

The data migration cost, equation **4-54**, is the cost of processing the migration action added to the cost of data transfer out from the storage server added to the re-ingest cost of newly generated objects (*ingC*$_{DMgrtDV}$).

$$DMgrtC = \left(\frac{DV}{DMgrtRP} \times * PC\right) + DToutC + ingC_{DMgrtDV}$$

4-54

The virtual appliance cost, equation **4-55**, is the cost of processing the virtual appliance added to the cost of data transfer out from the storage server.

$$VAC = (\frac{DV}{VARP} \times * PC) + DToutC$$

4-55

The above equations require the following input units:

a. Data Volume in Gigabytes (GB)

b. Time in Hours (hrs)

c. Retention Period(s) in Years (yr)

d. Costs in Euros (€)

e. Power Consumption in kilowatt hours (kWh)

4.3.7 Generation of Cost Assumptions

For the cost model to function, some inputs are pre-planned and designed to suit the behaviour of the preservation system. For both the models generated, private and public cloud, there are some assumptions that were made to permit cost estimation. These assumptions are editable and give users the flexibility to enter

their known operating conditions. However, there are some general assumptions valid for both models, whereas other assumptions are model dependent. These assumptions are listed as follows:

Cost Model Output Units:

i. All Cost/Prices are in Euros (€)

ii. All Data Volumes are in Gigabytes (GB)

iii. All Retention Periods are in Years (yr)

iv. Staff salary should include overheads and be quoted in Euros (€)

Public Cloud Cost Model Assumptions:

i. Compute Server choice

ii. Storage Cloud

iii. Existing Computer system infrastructure at the client to upload digital content from/to cloud

Private Cloud Cost Model Assumptions:
All construction works or cabling/trays costs are not included and it is assumed that a location for the data centre exists.

I. General Assumptions:

i. Electrical Power Cost: Average of European Industrial kWh cost

ii. Number of Working Days = 252 days

iii. Working Hours per day = 8 hours

iv. Choice of compute server is linked by the equivalent configuration to the choice in public cloud.

v. When an activity is running, the system is consuming the full load power rating.

vi. Idle power consumption is 30% of the full load, for compute servers only. Other hardware is constantly on full load.

II. Hardware Assumptions:

i. Servers:

a) Compute Server

b) Storage Server Model:

c) Servers' Chassis

d) Rack Enclosures

e) Storage Media All Hard Disk Drives, no tapes

ii. Networking:

a) Ethernet Modules

b) Switches

c) Distribution Switches

III. Software:

i. Cloud OS

ii. Storage Management Software

iii. Compute Server Operating System

IV. Other Assumptions:

i. Power System

a) Power Distribution System

b) Diesel Generators

c) UPS Model

ii. Cooling System

iii. Security System

iv. Fire Protection System

4.3.8 Single-Point Outputs from the cost model

Combining knowledge from WBS and CBS shows a high-level view of the cost generating activities. The costs of these activities are then calculated using the previously mentioned cost equations. The outputs of the cost model are the initial Investment cost, year one cost, annual running cost, total aggregate cost, total costs configuration horizon and total cost.

Initial investment defines the day zero costs to get the system up and running to a functional state; year one cost shows the cost performance of the LTDP system over the first 12 months; the annual running cost shows yearly cost performance.

The total cost is given in three tiers. Total aggregate cost displays the total cost for a specific aggregate of digital objects. The total cost configuration horizon displays the estimated cost over the period that the LTDP planners are confident of the successful performance of the configuration plan. ‘Total cost’ will display the estimated costs for the whole of the required duration.

The initial investment cost is computed as the sums of the initial data ingest cost, initial data transfer cost, initial data storage cost and software purchase/development costs. Year 1 cost is computed as the sum of the initial investment costs, the annual costs of the first year and annual staff cost.

The annual running cost is computed as the sum of the annual ingest cost, annual storage cost, annual access cost, annual data management cost and annual staff cost. This calculates the annual cost of the suggested configuration for all aggregates.

- If year is χ then add Transformation cost, where χ is the expected year of performing a preservation action

The total aggregate cost is computed as the sum of the ingest cost, storage cost, access cost, data management cost and staff cost for one specific aggregate, identified by its ID. This is calculated with the suggested configuration and data retention years selected by the cost model in the range indicated by the user.

- Staff costs can be calculated per aggregate by dividing the staff cost on the size percentage of the aggregate contribution to the total size of all aggregates.

- If year χ had occurred or had been expected to occur in the range, add Transformation cost, where χ is the expected year of performing a preservation action

The total costs configuration horizon is computed as the sum of the total aggregate costs for all aggregates for the number of years in the configuration horizon, adding the initial investment cost.

The total cost is computed as the sum of the total aggregate costs for all aggregates for the number of years up to the data retention period selected by the cost engine in the range specified by the user, adding the initial investment cost.

4.4 Chapter Summary

During this chapter, the method for reaching a single-point total cost estimation of the LTDP system was analysed in detail. Each step towards a cost estimation framework was designed with consideration to business sectors to achieve the most suitable technical results.

Understanding the LTDP lifecycle gives an initial comprehension of what should be researched as cost elements and LTDP activities. The account started with an exposition of the business sector differences which generated three cases, one for each business sector where the LTDP system is currently used. Thus, the view of cost elements was expanded and with analysis the key cost drivers were identified.

The clear flow of the work breakdown structure gave a comprehensive final view of the way in which the system is designed to perform in a manner to suit the client's requirements detected in the second process of the framework. The cost

breakdown structure was then constructed to reflect the main elements in the WBS that would generate cost along with the main key cost drivers.

From the CBS and WBS two ABC cost models were developed; one for each cloud computing deployment model and all the equations for both models were analysed as well as the way in which they related to other cost elements.

Guidelines were given showing how to design the assumptions for the cost modelling tool and how to relate them to previously discussed equations to achieve the set of outputs that the system is expected to provide. At this point, this research can provide the user with a cost model capable of estimating a single-point cost estimate.

To enable the user to have a better understanding of the cost estimation probability distribution and a more reliable cost estimates, a three-point estimate should be similarly designed to be complete this initial stage. The following chapter discusses how the discussed equations in this chapter can be expanded and how this expansion can generate a three-point estimate.

From the generated single and three points estimates a cost modelling process is developed, then the framework is realised.

5 QUANTIFYING UNCERTAINTIES AND OBSOLESCENCE ISSUES IN LONG-TERM DIGITAL PRESERVATION SYSTEMS

5.1 Introduction

In this chapter, the focus is on the conversion of single point cost estimates, generated from Chapter 4, to three-point cost estimates, then combining what's gained into an LTDP cost estimating process followed by a cost estimation framework. The framework will be composed of the high-level construction of the cost modelling process. This process will combine all the steps used to reach the combination of a single and three points estimate. This will ensure that the process and the framework have been constructed on the design of an actual cost model. This chapter is split into two sections, firstly moving the single to three-points estimate by combining the uncertainties and obsolescence issues to the single-point estimate. The second section of this chapter is constructing the cost estimation process and framework for LTDP.

5.2 Converting Single-Point Estimate to Three-Points Estimate

The conversion from single-point to three-points estimate is done by integrating the impact of uncertainties on cost elements in the cost drivers. The three-point estimates will be employed in a Monte Carlo Simulation to generate the probability distribution. The Monte Carlo simulation runs several iterations of the impact of uncertainties on cost drivers with their individual probability of occurrence; and the output is a probability distribution with three main values or a three-point estimate. Three-point cost estimates, as the most generic form of

quantitative uncertainty and risk analysis (Newton, 2009; MoD, 2007), provide "*numerical values to define a range of possible outcomes*" (MOD, 2007). The numerical values are displayed with respect to their probability distribution.

Instead of providing the cost estimator with a single number for any estimate, three point estimates can be returned in the form of Minimum, Most Likely and Maximum cost values (Roy, 2011; Erkoyuncu, 2009; Newton, 2009; MoD, 2007). The minimum cost values represent the best-case scenario that estimators set, where cost values are favourable to the estimator. The most likely returns cost values that have the highest probability of occurring. The maximum cost values represent the worst-case scenario that estimators set, ignoring only highly unexpected events, such as natural disasters (Roy, 2011; MoD, 2007).

In the following sections of this chapter, the process of integrating in cost drivers the impact of uncertainties on long-term digital preservation systems is described, including obsolescence.

5.2.1 Uncertainty Cost Estimation

This section discusses uncertainties, in general, and uncertainties in LTDP. Defining uncertainties and how to identify, score and evaluate them in an LTDP system is shown and investigated in some detail. Uncertainty in general means the lack of certainty or the lack of precise description or knowledge of something or an outcome. The area of uncertainty in research is vague and grasping the concepts of uncertainty in a certain discipline (digital preservation, in this case) requires a thorough and rigorous understanding of the processes and/or actions under investigation.

Understanding uncertainty started by trying to develop mitigation strategies to avoid or reduce the impact of uncertainties or processes or actions (Shehab, 2013; Chuku, 2012; Erkoyuncu, 2011). While mitigation strategy development is crucial, so is understanding the definition of uncertainty from the right perspective while taking care to attain the targets. Walker et al. (2003) defines uncertainty as "any deviation from the unachievable ideal of completely deterministic knowledge of the relevant system", while McManus and Hastings (2004) link the definition of uncertainty with vagueness and doubtfulness. Erkoyuncu (2011) defines uncertainty as "stochastic behaviour of any physical phenomenon that causes the indefiniteness of outcomes", linking it also with "lack of knowledge". In the case of cost estimation, this definition can be considered closest to the target perspective of the present research.

McManus and Hastings (2004) divide uncertainty, once realised, into two kinds: an uncertainty with a negative outcome is a risk and an uncertainty with a positive outcome is an opportunity. Erkoyuncu (2011) differentiates between uncertainty and risk. He describes risk as "a case of uncertainty" that can "have a negative effect". After identifying uncertainties, mitigating risks and exploiting opportunities are likely to prove most beneficial.

Therefore, uncertainties in cost estimation for long-term digital preservation may be defined as

"the random incidents …[whose] occurrence cannot be predicted in advance, generating a deviation from expectation. These incidents can happen due to different factors that will affect an LTDP system, preserved data within the system

or in managing the system. These incidents will impact the functionality expected from the LTDP system in a way that will make the estimated cost less accurate and will add vagueness to future predictions”

Over time, researchers have developed an understanding of the nature of uncertainty. Uncertainty may depend on randomness, for example, a throw of the dice, known as an aleatory uncertainty. This cannot be avoided or even predicted by bringing in more information or knowledge. If, however, the uncertainty relates to a limitation of knowledge, it is known as an epistemic uncertainty. This type may have less of an impact when more understanding and information are brought in (Erkoyuncu, 2011; Thunnissen, 2005; Smith, 2002). Figure **5-1** shows the different kinds of uncertainty.

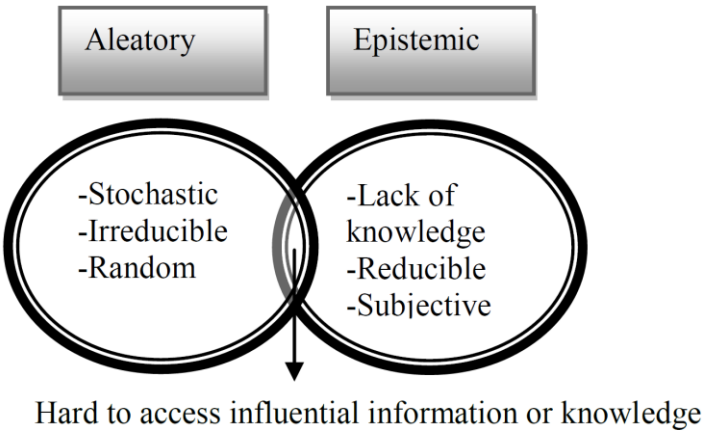


Figure 5-1 The two Kinds of Uncertainty (Erkoyuncu, 2011)

To integrate uncertainty with the cost estimate, NATO (2007) has set up a procedure. The following six steps enable the uncertainty factors to be merged with the relevant cost elements (Roy, 2011; NATO 2007):

1. Identify all the affected cost drivers
2. Generate a single point cost estimate

3. Add the influence of uncertainty to the single point estimate of the cost driver, by adding the probability distribution
4. Select the most suitable type of distribution
5. Estimate the parameters of the distributions, i.e. the minimum, the most likely and the maximum (three-point estimate)
6. Use the three-point estimate as an input to a Monte Carlo simulation to realise the effect of uncertainties on each cost driver

Following these six steps will produce a three-point estimate with a probability distribution. In a Monte Carlo simulation, these values can be run randomly over many iterations and combine them in the resultant probability distribution. From the above procedure, it is clear that reaching a reliable output of collected data is crucial. In the following sections, the methodology for obtaining a three-point estimate is outlined. Next, an uncertainty identification process is described, to show how to extract information about possible uncertainties if it is not already known, and how to find the size of the impact of uncertainties residing in the actions of the LTDP system. Finally, obsolescence as an uncertain factor is discussed, together with the size of its impact on cost drivers. Obsolescence was by its nature a major topic in the present research project, due to the vulnerability of all its elements to technological change.

5.2.2 Methodology

To reach the targets of this research and capture the impact of uncertainties on cost, the procedure of integrating uncertainties on cost drivers was followed, with a small addition. The work of Phase 1 was to collect data from 27 interviewees from a diverse range of business sectors (see Table **5-1**), to broaden the sources

of data. The difference between the procedures of phase 1 and those of NATO (2007) is that phase 1 finds the cost drivers and the single-point estimate (see Chapter 4) and begins by identifying what uncertainties could exist in LTDP systems. In Phase 1, experts are selected based on their area of expertise. They had to be working with LTDP systems or part of that system for at least 1 year. They had to cover all areas from the WBS of the LTDP shown in Figure 4-3 or any of its sub WBSs.

Phase 2 contains an analysis of the interviewees' answers and extracts an uncertainty impact score for each uncertainty, followed by the application of the impact scores to the cost drivers.

Table 5-1 Interviewees for Uncertainty Identification Process

No.	Company	Job role	Years of Experience
1	Hewlett Packard	Component Engineer	25
2	BAE Systems	Principal Scientist	10
3	Science & Technology Facilities Council (STFC)	Project Manager	25
4	Network for Earthquake Engineering Simulation (NEES)	Data Curator	4
5	Custodix	Project Manager	2
6	Digital Curation Centre (DCC)	Associate Director	7
7	IBM	Managing Consultant	15
8	London School of Hygiene & Tropical Medicine	Research Data Project Manager	10
9	University of Manchester	Researcher	3
10	Kings College London	Digital Archivist	3
11	University of Birmingham	Institutional Repository Manager	4
12	Santa Fe Institute	Librarian	10
13	Phonogrammarchiv	Video Technician	10
14	Systems Research & Applications (SRA) International	Senior Cyber Security Engineer	12
15	University of California	Director	25
16	University of Tasmania	Emeritus Professor	30

17	Brooklyn Historical Society	Director of Library Archives	12
18	Science & Technology Facilities Council (STFC)	Head, Operations	10
19	Duke University Archives	Electric Records Archivist	5
20	University of Exeter	Senior Technical Manager	15
21	De Montfort University	Repository Officer	4
22	University of Maryland	Professor	15
23	STFC	e-Science	18
24	Tessella	Digital Archiving Solutions Development	22
25	Tessella	Digital Preservation Technologies Development	10
26	Custodix	Digital Security Development	11
27	University of Porto	Economic Performance	4

To find the impact of uncertainty issues on the cost drivers of the LTDP system, the interviewees were asked to rate what they perceived to be uncertain in each preservation activity. Afterwards each value on a scale from 1 to 9 represented the impact of each uncertainty on the cost drivers. An example of the scaling of impacts on uncertainties is shown in Table **5-2**.

Table 5-2 Impact of Uncertainty Issues on Cost of LTDP Systems: An Example

Uncertainty Issue	**Impact on Cost Drivers**								
	No impact	Low Impact but not negligible			Balanced	High Impact but not Extreme			Extreme Impact
Hardware Cost Negotiation	1	2	3	4	5	6	7	8	9

Next, each uncertainty was linked with the frequency or probability of its occurrence to meet the requirements for a Monte Carlo simulation. An example of the probability of occurrence for uncertainties questions is shown in Table **5-3**.

Table 5-3 Probability of Occurrence of Uncertainties: An Example

Uncertainty Issue	Probability										
Hardware Cost Negotiation	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1

At the end of the interviews the interviewees attended an open discussion to give them a chance to agree on each value for the impact or probability of uncertainty factors on cost elements. This was meant to reduce possible subjectivity and bias in the interviewees. Reducing bias means reducing the impact of innate optimism or pessimism on the opinions expressed in the interviews; viewing a variety of results and having an open discussion based on the Delphi method, questionnaire appendix A.3 and A.6, can lead to consensus on a value or a smaller range of values; thus, improving understanding.

5.2.2.1 Defining Uncertainty Categories in Long-Term Digital Preservation

Before understanding the impact of uncertainties on cost elements, the sources of uncertainty were discussed and elaborated, to aid the discovery of more uncertain events and to aid the experts in determining the correct scoring for each uncertainty.

Starting the process of identifying the sources of uncertainty resulted in discovering five main categories, as shown in Figure **5-2**.

- Economic uncertainties are the uncertainty issues arising from financial conditions, such as inflation, power costs, the availability of company resources, etc.
- Technological uncertainties depend mainly on the progress of technology and the technological capability of the company. These

uncertainties include obsolescence issues, hardware and software upgrades, and engineering capabilities, the correct choice of hardware/software, the correct installation and commissioning of hardware/software and the complexity of the data.

- Business related uncertainties all stem from the management of the company/organisation. They include security, preservation selection/planning, organisational change, stakeholders' expectations, customer relationships, supplier relationships and relationships with data owners.
- Uncertainties generated from the relationship with the government are called regulatory, because the system is directed by local regulations, such as government incentives.
- Physical uncertainties are usually related to natural disasters or equipment failure; they may result from a storage failure, hardware failure, infrastructure scalability issues, inadequate infrastructure availability and of course natural disasters.

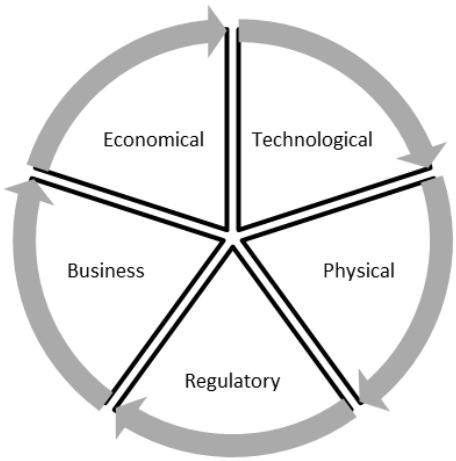


Figure 5-2 LTDP Uncertainty Categories

These categories can be very useful in learning how to quantify the impact of uncertainties on cost elements.

5.2.2.2 Uncertainty Identification Process

The uncertainty identification process, shown in Figure **5-3**, fits in the cost estimation framework because it identifies the nature of the uncertainty.

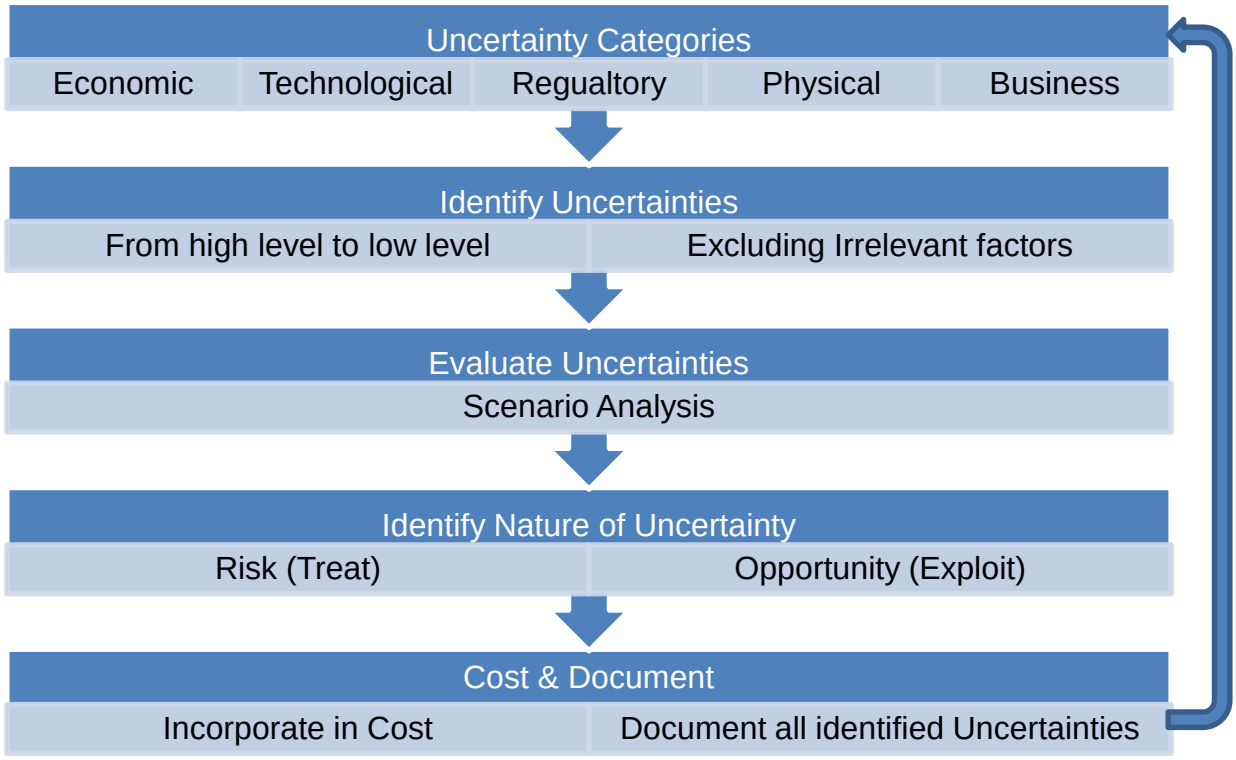


Figure 5-3 The Uncertainty Identification Process

The nature of the uncertainty as outlined in the introduction to this chapter could be risk, with a negative effect on cost, or opportunity, with a positive effect on cost. These uncertainties are linked to cost equations and to specific cost elements. The cost elements are the ones that will be affected by these specific uncertainties. The preservation uncertainties, along with the impact of obsolescence issues, are quantified by the appropriate cost metrics.

Each uncertainty category affects the cost equations differently. Some of the effect is negative, i.e. entailing risk, thus increasing the cost by a factor: others affect the cost positively, i.e. offering an opportunity, thus reducing the cost, also by a factor. The cost of uncertainty factors was quantified in the interviews with digital preservation experts.

The effect and combination of uncertainties and obsolescence issues on the preservation cost metrics generates a three-point cost estimate. The original cost model output is the most likely value and the experts provided the expected minimum and maximum value of the cost estimate considering their experience, also identifying which uncertainty factor would affect which preservation cost metrics.

Table **5-4** illustrates a list of different cost elements along with the uncertainty factors, which was analysed from questionnaire in appendix A.4. A negative percentage is a factor that should be subtracted from the related cost elements, i.e. opportunity, and a positive percentage is a factor that will be added to the related cost elements, i.e. risk (Min= Minimum Cost, Mo= Most Likely Cost and Max= Maximum Cost):

Table 5-4 Impact of Uncertainties on the Corresponding Cost Elements

No.	Title (Cost Elements)	Minimum	Most Likely	Maximum	Occurrence Probability
a.	**Infrastructure Acquisition Costs**				
i.	Hardware Cost	-70%	-40%	0%	0.8
ii.	Software Cost	-30%	0%	+30%	0.5
iii.	Other Cost	-10%	+20%	+50%	0.4
b.	**Implementation Costs**				
i.	Installation Cost	-30%	0%	+30%	0.2
ii.	Configuration Cost	0%	+20%	+50%	0.6
iii.	Commissioning Cost	-10%	+20%	+50%	0.2

c.	**Running Costs**				
i.	Idle Power Consumption Cost	-50%	-20%	+10%	0.8
ii.	Internet Connection Cost	-50%	-10%	+30%	0.9
iii.	Maintenance Cost	-10%	+30%	+70%	0.8
iv.	Software Licences Cost	-10%	0%	+10%	0.8
d.	**Preservation Activities Costs (*Private cloud based system*)**				
i.	Full Load Power Consumption Cost	-20%	+10%	+40%	0.7
ii.	Staff Cost	-10%	+20%	+50%	0.9
e.	**Processing Costs (*Public cloud based system*)**				
i.	Public Cloud Providers' Pricing	-40%	-10%	+20%	0.4
ii.	Operating System Cost	-10%	0%	+10%	0.4
iii.	Utilisation	-10%	0%	+10%	0.8
iv.	Memory Available	-10%	0%	+10%	0.7
f.	**Ingest Costs**	-20%	0%	+20%	0.8
g.	**Storage Costs**				
i.	Data Transfer Cost	-40%	-10%	+20%	0.9
ii.	Storage Volume Cost	0%	+10%	+20%	0.4
h.	**Data Management Costs**	-20%	0%	+20%	0.7
i.	**Access Costs**				
i.	Information Package Retrieval Cost	-20%	0%	+20%	0.4
ii.	Data Protection Cost	-50%	0%	+50%	0.5
j.	**Transformation Costs**	-20%	0%	+20%	0.3

All the above values were applied to the corresponding equations; they gave a final total cost ranging from a minimum value to a maximum value. For example, if the Transformation Costs is x, then:

$$\boldsymbol{Min = x - 20\%, Mo = x + 0\%, and Max = x + 20\%} \qquad \textbf{(5-1)}$$

5.2.3 Obsolescence and Long-Term Digital Preservation

Shehab, et al. (2013) identified obsolescence as the major technological uncertainty of any LTDP system, while Xue, et al. (2011) considered obsolescence a major challenge in cost modelling for LTDP. In the generic understanding of engineering, obsolescence usually concerns components or technology. Obsolescence in manufacturing is defined as the state when "it is no

longer manufactured or supported by its original manufacturer or a third party. This happens either because demand has dropped to low enough levels that manufacturers choose not to continue to make it, or because the materials or technologies necessary to produce it are no longer available" (Sandborn, 2007a; Singh, et al. 2006).

Within LTDP, obsolescence has been defined as

"*digital information [that] is still at hand, but not readable because the media's reader (the hardware or software) is no longer available. The main issue is that software and hardware technology becomes rapidly obsolescent. Storage media become obsolete as do devices capable of reading such media; and old formats and standards give way to newer formats and standards*" (Neervens, 2009; Waters, et al. 1996).

Many libraries and heritage organisations are already doing their utmost to protect human history and have embarked on digital preservation as a core business tool. Understanding the costs of digital preservation activities is very important for organisations that preserve human culture, science and history, because the stored information could be priceless; thus, the value of the information value diminishes the importance of the costs incurred to preserve it. However, for the business sector the understanding of costs incurred from carrying out LTDP activities becomes more immediately important. Cost understanding becomes a support tool for decision makers who must decide whether or not to commit to a strategy of spending funds for a considerable number of years to preserve their company's sensitive information.

5.2.3.1 Methodology

The main aim in this section of the research is to define what obsolescence issues will impact on an LTDP system and how to capture their impact on cost. The methodology is composed of three main phases.

Phase 1 focuses on collecting knowledge from the literature and experts in the LTDP field. This phase involved experts from all 13 partners of ENSURE in addition to experts from the World Health Organisation, Alliance for Permanent Access, ACCENTURE and many more. Experts were interviewed in their area of expertise. For experts managing LTDP systems the questions concerned high level and strategic issues, while experts with hands-on technical experience faced questions on low level technical issues.

The output of these interviews, along with the understanding and findings from the literature, was analysed as phase two. A comparative analysis was carried out that resulted in a detailed taxonomy of obsolescence issues. shown in Figure **5-4**.

Finally, in phase three, obsolescence mitigation strategies were collected from experts and cost equations were developed. A workshop with 17 participants from the LTDP domain was held to identify the direct impact factor of each element of obsolescence on its corresponding cost metric.

5.2.3.2 Quantifying Obsolescence in LTDP systems

To find the impact of obsolescence issues on cost, questions were put to the experts to answer on a scale from 1 to 9. Each value represented the strength of impact of obsolescence issues on cost. The impact scale of obsolescence issues used in the present research is shown in Table **5-5**.

Table 5-5 Impact scale of Obsolescence issues

Obsolescence Issues	Impact on Cost Drivers								
	No impact	Low Impact but not negligible			Balanced	High Impact but not Extreme			Extreme Impact
Hardware	1	2	3	4	5	6	7	8	9
Software	1	2	3	4	5	6	7	8	9
Human Skills	1	2	3	4	5	6	7	8	9
Preservation Strategies	1	2	3	4	5	6	7	8	9

Questions for this workshop is in appendix A.5. The scale chosen was based on the Likert scale (Likert, 1932). The scale ranged from no impact, minimum, and balanced, to high and extreme impact. This was chosen to ensure the least possible bias and to provide a midpoint for experts who wanted to indicate a neutral point in the scale.

Experts attending the workshop came from several companies and had a range of years of experience, thus ensuring diversity in the sample and the reduction of bias. A summary of the length of experience and background of those who attended is given in Table **5-6**.

Table 5-6 Experts attending the LTDP Workshop

No.	Company	Job role	Years of Experience
1	CJ-IMC	Independent Consultant	10
2	Tell Berlin	Research Data Coordinator	3
3	NLS	Digital Preservation officer	13
4	Digital repository of Ireland	Digital Archivist	1
5	DANS – Data Archiving and Networked Services	Data Manager	6
6	FRD	N/A	10

7	UC3M	Assistant Prof. Digital Libraries Masters Course	3
8	An Educational Institution	Librarian	1
9	Digital Preservation Coalition	Executive Director	15
10	CINECA	Software Integrator	2
11	Charles University in Prague	System Administrator	5
12	Digital Preservation Coalition	Senior Project Officer	7
13	Digital Preservation Coalition	Senior Project Officer	1
14	KEEP Solutions	Innovation Director	7
15	Fuse – Institute Berlin	Research Developer	3
16	University of Leeds	Digital Content & Repositories Manager	13
17	VIAA	Digital Archivist	1

5.2.3.3 Taxonomy of Obsolescence

Obsolescence issues in LTDP systems can be found in four main categories: Hardware, Software, Human Skills and the Preservation Plan.

Figure **5-4** is a diagram of the full taxonomy of obsolescence issues in the LTDP system. Each category has subcategories that show the exact source of cost.

a. <u>Hardware Obsolescence</u>

Both management and technical experts agree that the most important obsolescence issue is that of mitigating hardware obsolescence. Digital preservation practitioners usually update all their hardware after 3 to 5 years in service, as a best practice precaution. This is directly linked with the warranty provided by the hardware manufacturers.

The experts agreed on the life range, but some insisted that they never ran the hardware above the provided warranty, while others confirmed they could run it to as much as 2 years above the provided warranty. Generically this hardware

includes the computing system or part of it, the peripherals and the complete storage media system, both readers and media.

LTDP practitioners are often worried to find that mitigating hardware obsolescence costs more than is justified by the impact it has on preservation systems. This suggests that hardware obsolescence is technically not as challenging as other obsolescence issues.

b. Software Obsolescence

Sandborn, et al. (2007a, 2007b) highlights the causes of software obsolescence for Commercial-Off-The-Shelf (COTS) software packages as follows:

a. Functional Obsolescence: where any change to the computer system affects the expected functionality of the software.

b. Technological Obsolescence:

    i. Software no longer sold by the original supplier (end of sale).

    ii. Inability to renew or expand licence agreement (legal).

    iii. Original Supplier no longer supports the software (end of support)

c. Logistical Obsolescence: media obsolescence, formatting, degradation which limits or terminates the accessing software.

Experts in the workshop showed a high level of awareness and a similar level of worry about software obsolescence. The reason behind this is that it is unknown

which software application or file format will become obsolete or when it will do so.

It may not generate much cost but it is beyond doubt considered a major obsolescence issue. The uncertainty is too high and even now no LTDP system is considered successful unless it can handle the impact of obsolescence. However, they agree that it should not be treated as a low-cost impact issue. Not only is the file format or the software application endangered, but also any plug-ins to the main software or even the operating system. Furthermore, the risky prospect of a software provider dropping the backward compatibility strategy was found to be highly likely, resulting in the loss of the required software.

c. Human Skills Obsolescence

Human skills obsolescence is not usually noticed, since organisations tend to provide regular training to their employees. However, workforce skills do become obsolete over time, even without the loss of specific human resources (Sandborn, et al. 2012).

As soon as the question was introduced, experts recognised Human Skills Obsolescence as a major category of LTDP obsolescence. They specifically saw that a loss of skills in using particular software packages can be common.

The loss of ability to handle a software package or the limited use of applications can be managed; but the major risk is that an employee will fail to extract useful information from a digital content despite its successful technological preservation.

d. Preservation Plan Obsolescence

The final obsolescence issue, which was not easy to discover, was the obsolescence of the Preservation Plan; in this case, the company's plans and strategies are not going to be successful in meeting the company's requirements. The whole of the preservation system is at risk if it is hit by the consequences of this issue.

New file formats could be lost; essential software might be missed from an old company strategy or other obsolescence issues in the company plan may even be mishandled.

In the ENSURE preservation system, cost values are provided in two different forms. The first shows the total cost over the preservation horizon, which is the period that a certain preservation plan is expected to last. The second shows the total cost over the total required preservation period. This defines a set time frame in which to review the company preservation strategy.

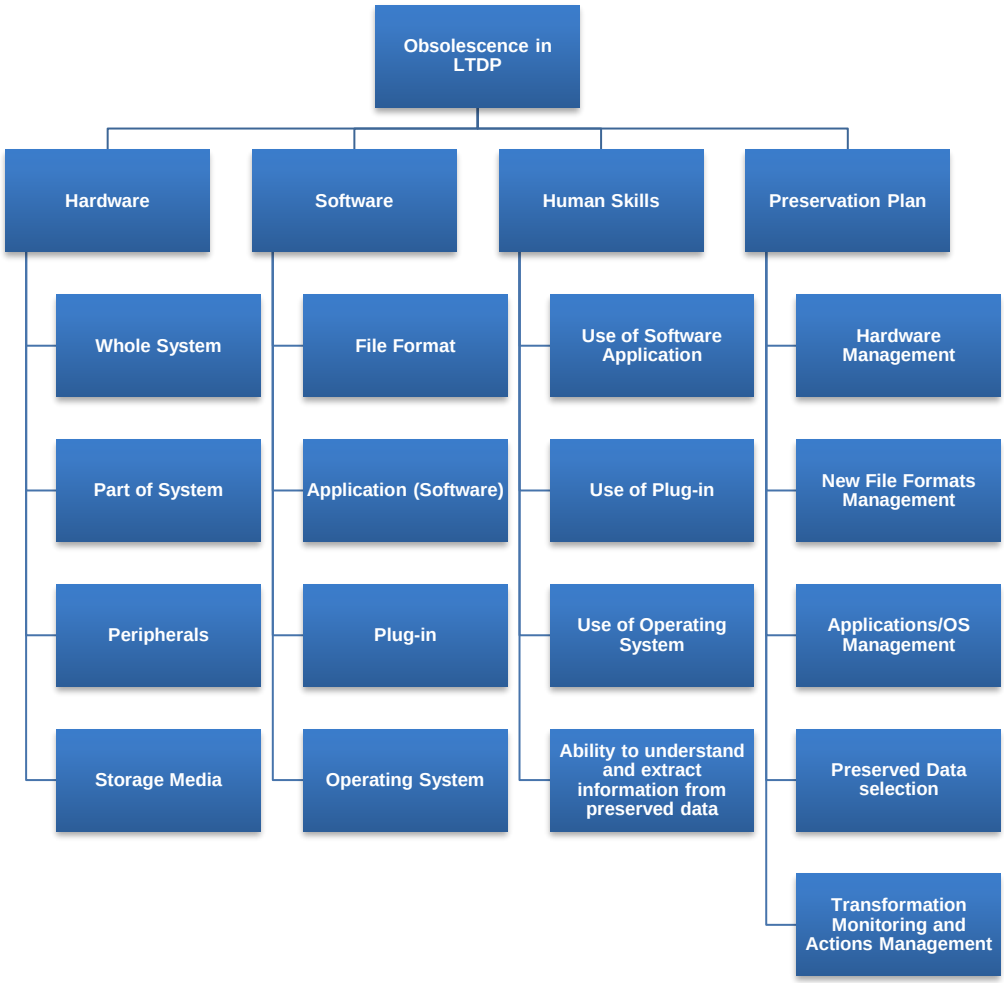


Figure 5-4 Obsolescence in the LTDP Taxonomy

5.2.3.4 Obsolescence Cost

After generating the obsolescence taxonomy, cost equations for obsolescence mitigation were developed. Each main category is quantified by its own equation. The developed equations here derive from a private cloud based LTDP system.

- **Hardware obsolescence:**

$$Hardware\ Obsolescence\ Mitigation\ Cost \left(\begin{matrix} Repeated\ Cost\ 3\ to\ 5\ year\ according\ to \\ company\ strategy \end{matrix}\right) = Compute\ Servers\ Cost + Storage\ Servers\ Cost + Networking\ Equipment\ Cost + Server\ Enclosures\ Cost + Storage\ Media\ Cost \quad (5\text{-}2)$$

- **Software obsolescence**

$$Software\ Obsolescence\ Mitigation\ Cost\ \begin{pmatrix} licences\ repeated\ according\ to \\ each\ licence\ policy \end{pmatrix} = Software\ Licences\ Cost + Migration\ Action\ Cost$$

(5-3)

$$Software\ Licences\ Cost = Operating\ System\ Licence\ Cost + annual\ Storage\ Management\ Software\ Cost + other\ software\ Cost$$

(5-4)

$$Migration\ Action\ Cost = Data\ Volume \times IT\ System\ full\ load\ Power\ Consumption\ Cost \times IT\ system\ Processing\ Rate \times Staff\ Cost\ per\ hour$$

(5-5)

- **Human Skill Obsolescence**

$$Human\ Skill\ Obsolescence\ Mitigation\ Cost = Staff\ Cost\ per\ Hour \times Number\ of\ Staff \times Training\ Course\ Cost \times Number\ of\ Courses\ per\ Year$$

(5-6)

- **Preservation Plan Obsolescence (Re-planning)**

$$Preservation\ Strategy\ Obsolescence\ Mitigation\ Cost\ (Added\ at\ the\ End\ of\ the\ Plan\ Horizon) = (Number\ of\ Staff \times Time\ Required \times Staff\ Cost\ per\ Hour) \pm (Reduction\ or\ increase\ in\ data\ volume\ cost\ OR\ reduction\ or\ increase\ in\ processing\ power\ requirements)$$

(5-7)

To realise the cost of these obsolescence issues, their individual cost values from the above equations were multiplied by the figures for their minimum, most likely and maximum impact. The resultant totals were added to the corresponding LTDP cost values.

5.2.3.5 Obsolescence Cost Impact Factors

As part of the ENSURE cost report requirement, costs must be presented in a three-point estimate format, the user is provided with an expected minimum, expected most likely and expected maximum cost. To calculate the three-point

estimate for the impact of obsolescence issues, the minimum, most likely and maximum impact factors were ascertained in a workshop with experts in the LTDP field. Obsolescence analysis was carried out in a workshop of 17 experts, with experience in LTDP. The participants were asked to suggest a weight for the actual or probable impact on cost. The weight of impact on cost was measured on a scale from 1 to 9 with 1 step increments. These odd steps were intended to reduce bias and contained a mid-value for candidates who wanted to provide a neutral answer.

The Most Likely value is calculated according to the weight of the contribution of each single answer to the whole question. Therefore, the total sum of all values, each multiplied by the number of votes achieved, was then divided by the total number of participants, as shown in Equation (**5-8**.

$$\frac{\sum_1^9 \boldsymbol{s}\times\boldsymbol{v}}{\boldsymbol{n}}$$

(5-8)

s =scale value selected. v =number of votes. n = total number of participants

The impact percentages calculated for these obsolescence issues are always positive in value, meaning that they add to the final cost of preservation, because they are always considered to be a risk. These impact factors are then multiplied by the final cost result of each obsolescence issue and finally added to the total cost of preservation. For each of the obsolescence issues the Maximum (Max), Minimum (Min) and Most Likely (MO) impact is shown in Tables 7, 9, 11 and 13. The following pie charts show each impact factor, 1 to 9, and the percentage of

candidates who voted for each value, similarly to frequency of occurrence, from 0 to 1, and the percentage of votes.

Table **5-7** shows the voting result for the hardware obsolescence issue as a whole, from the interviewees' point of view. However, these values are not used in the cost model, since they are high level and not specific enough. They were collected only to familiarise the experts with the categories of obsolescence.

Table 5-7 Collective Impact Factor for All Hardware Obsolescence Issues

Obsolescence Issue: Hardware			
Impact			Expected Frequency of Occurrence
Min	**MO**	**Max**	**MO**
3	6	9	0.63

Table **5-8** shows all the hardware obsolescence sub-category votes for impact on cost and probability. The voting shows how the experts see hardware obsolescence as a major event, since all of them agreed that all its sub categories had a very high cost impact. This is due to the high cost of procuring replacement hardware and the unexpected phasing of technologies.

Table 5-8 Obsolescence Hardware Sub-Issue: Scores

Obsolescence sub-Issue: Hardware – Whole System			
Impact			Frequency of Occurrence
Min	MO	Max	MO
2	5.75	9	0.46
Obsolescence sub-Issue: Hardware – Part of System			
Impact			Frequency of Occurrence
Min	MO	Max	MO
3	6.06	9	0.69
Obsolescence sub-Issue: Hardware – Peripherals			

Impact			Frequency of Occurrence
Min	MO	Max	MO
1	4.75	9	0.71
Obsolescence sub-Issue: Hardware – Storage Media			
Impact			Frequency of Occurrence
Min	MO	Max	MO
3	6.81	9	0.68

General software obsolescence votes are shown in Table **5-9**. The understanding of this category among the experts is similar to their understanding of hardware obsolescence. Undeniably, all the experts agreed that it should have a maximum impact of 8 on the Likert scale. This is due to the high cost of mitigating this form of obsolescence. The high salaries of software engineers and programmers, spread of proprietary software that is common in the long-term digital preservation field and the specific scope of businesses in the sectors, made it very easy to rank software obsolescence amongst the most serious impact factors.

The difference between the hardware and software obsolescence votes is clear from the probability of occurrence of either; experts decided that software obsolescence was the more likely of the two to occur and impact a data set that is being preserved for a given time.

Table 5-9 Collective Impact Factor for All Software Obsolescence Issues

Obsolescence Issue: Software			
Impact			Expected Frequency of Occurrence
Min	MO	Max	MO
3	6.47	8	0.76

Looking at the individual obsolescence sub category votes in Table **5-10**, one can identify the level of concern in the LTDP community about software

obsolescence. The lowest probability category is that of operating systems, though it still has the highest impact. A long discussion was needed to agree upon a value; most of the controversy was occasioned by the obsolescence of Windows XP® as an operating system, for which Microsoft withdrew its support on 8th April 2014 after it had been used stably by many in the community for 12 years (Microsoft, 2014).

Following the OS obsolescence predicament, other sub categories show higher probability ratios with similar impact factors. This implies the fragility of data preserved because of a company's dependence on its own programming language or some other proprietary software package. The problem was simply accepted among the experts as follows: "*Yes, any software can become obsolete very quickly*", "*A smarter line of code can change a lot*" and "*It happens every day*".

Table 5-10 Obsolescence Software Sub-Issue: Scores

Obsolescence sub-Issue: Software – File Formats			
Impact			Frequency of Occurrence
Min	MO	Max	MO
3	5.63	8	0.7
Obsolescence sub-Issue: Software – Applications			
Impact			Frequency of Occurrence
Min	MO	Max	MO
2	5.81	9	0.71
Obsolescence sub-Issue: Software – Plug-ins			
Impact			Frequency of Occurrence
Min	MO	Max	MO
2	4.33	8	0.73
Obsolescence sub-Issue: Software – OSs			
Impact			Frequency of Occurrence
Min	MO	Max	MO
2	5	9	0.55

Human skills as an obsolescence category generated debate between experts when introduced by the researcher as a possible obsolescence issue. Some would not recognise it as an issue at all, while others agreed that it was a problem for data users more than anyone else. The problem appears when the data user retrieves a preserved data set but has lost or never had the knowledge to interact with it. The whole idea behind LTDP is to provide users with valuable data, so if a user fails to interact with the preserved data, the result is equivalent to a loss of data.

At the end of the discussion the interviewed experts agreed that the obsolescence of Human Skills was less likely than Hardware and Software obsolescence, but nevertheless agreed that when it occurs it has a similarly high impact on cost. Loss of Human Skills can always be mitigated by training and retraining the staff who have access rights.

Table 5-11 Collective Impact Factor for All Human Skills Obsolescence Issues

Obsolescence Issue: Human Skill			
Impact			Expected Frequency of Occurrence
Min	**MO**	**Max**	**MO**
3	6.11	8	0.57

The main concern came from the sub category “Extracting Information from Preserved Files”. It is shown in

Table **5-12** that its score was very high regarding impact on cost and it was the one with the highest frequency of occurrence. Applications, plug-ins and OSs can be handled by most LTDP experts since they are all external factors, but

understanding internal data may depend on a single employee or a specific task in the firm that no longer exists.

Table 5-12 Obsolescence Human Skills Sub-Issue: Scores

Obsolescence sub-Issue: Human Skills – Use of SW Applications			
Impact			Frequency of Occurrence
Min	MO	Max	MO
2	4.94	9	0.57
Obsolescence sub-Issue: Human Skills – Use of Plug-ins			
Impact			Frequency of Occurrence
Min	MO	Max	MO
1	4.13	7	0.57
Obsolescence sub-Issue: Human Skills – Extracting information from Pres. files			
Impact			Frequency of Occurrence
Min	MO	Max	MO
2	5.75	9	0.69
Obsolescence sub-Issue: Human Skills – Use of OSes			
Impact			Frequency of Occurrence
Min	MO	Max	MO
1	4	8	0.41

The final obsolescence issue was the obsolescence of the preservation strategy. The votes suggested that this obsolescence issue had the least impact and the least probability, as shown in Table **5-13**. Experts expressed their usual readiness with this kind of obsolescence to meet the targets set by the preservation planners. A solution can be prepared fairly easily if a new strategy is employed. In some cases, this might be very difficult, but it was voted the most predictable type of obsolescence.

Table 5-13 Collective Impact Factor for All the Preservation Plan Obsolescence Issues

Obsolescence Issue: Preservation Strategy			
Impact			**Expected Frequency of Occurrence**
Min	MO	Max	MO
3	**5.24**	**7**	**0.46**

In detail, formulating a strategy for a software application and choosing the data to be preserved scored the highest impact on cost, as shown in Table **5-14**. Selecting the data to be preserved is a very interesting sub-category; since the entire cost comes from keeping these data sets viable and accessible. Therefore, the experts all agreed on the seriousness of its impact, especially if a poor choice was made at the strategic stage.

Table 5-14 Obsolescence Preservation Strategy Sub-Issue: Scores

Obsolescence sub-Issue: Preservation Strategy – HW System Management			
Impact			Frequency of Occurrence
Min	MO	Max	MO
2	5.44	7	0.49
Obsolescence sub-Issue: Preservation Strategy – SW Applications/System Management			
Impact			Frequency of Occurrence
Min	MO	Max	MO
2	5.75	8	0.61
Obsolescence sub-Issue: Preservation Strategy – New Formats Management			
Impact			Frequency of Occurrence
Min	MO	Max	MO
2	5.07	7	0.59
Obsolescence sub-Issue: Preservation Strategy – Preserved Data/Info Selection			
Impact			Frequency of Occurrence
Min	MO	Max	MO
2	5.25	8	0.52
Obsolescence sub-Issue: Preservation Strategy – Transformation Monitoring/Actions Mgt.			
Impact			Frequency of Occurrence
Min	MO	Max	MO
3	5.63	7	0.61

5.3 LTDP Cost Modelling Process

Based on reaching a single and three-points estimates, now a cost modelling process can be deduced. What the researcher did is:

1. Understand Digital Preservation Current Best-Practice
    a. Understand the LTDP lifecycle: which when combined with cost drivers enables the development of a work breakdown structures
    b. Highlight differences in sectors requirements: will highlight important cost metrics for business sectors
    c. Find key cost drivers: thus, enabling WBS construction
2. Analyse Cost Data
    a. Breakdown work into a structure (WBS): finds relationships between effort made within an LTDP system
    b. Breakdown cost into a structure (CBS): finds relationships between effort made and cost incurred due to that effort
    c. Generate A single point estimate equations, assumptions and rules: generating the equations and rules at this stage will enable a clear implementation of uncertainties and obsolescence issues impact in the following steps. This will clarify cost elements with each cost driver, thus making it simpler to identify which impact will affect what part of the equation.
    d. Identify uncertainties impact on cost drivers:
        i. Utilise uncertainties identification process
        ii. Generate an impact factor for each identified uncertainty
    e. Identify obsolescence impact on cost drivers:
        i. Use obsolescence taxonomy to generate appropriate mitigation strategies, thus mitigation costs
        ii. Generate an impact factor for each obsolescence issue

3. Cost estimation and results

    a. Combine single-point equations with the uncertainties and obsolescence issues impact on drivers and impact between issues

    b. Generation of estimation tool: a spreadsheet based tool can be fast at an initial stage

    c. Simulating impact factors probabilities on cost drivers: Monte Carlo, triangular distribution

    d. Generate a three-points estimate

Full estimation process can be seen in Figure 5-5. In the diagram, black arrows mean forward progress and red arrows refer to processes that can feedback to previous or other coming processes. If followed as shown in Chapters 4 and 5, an LTDP user can generate their own business specific cost model for LTDP.

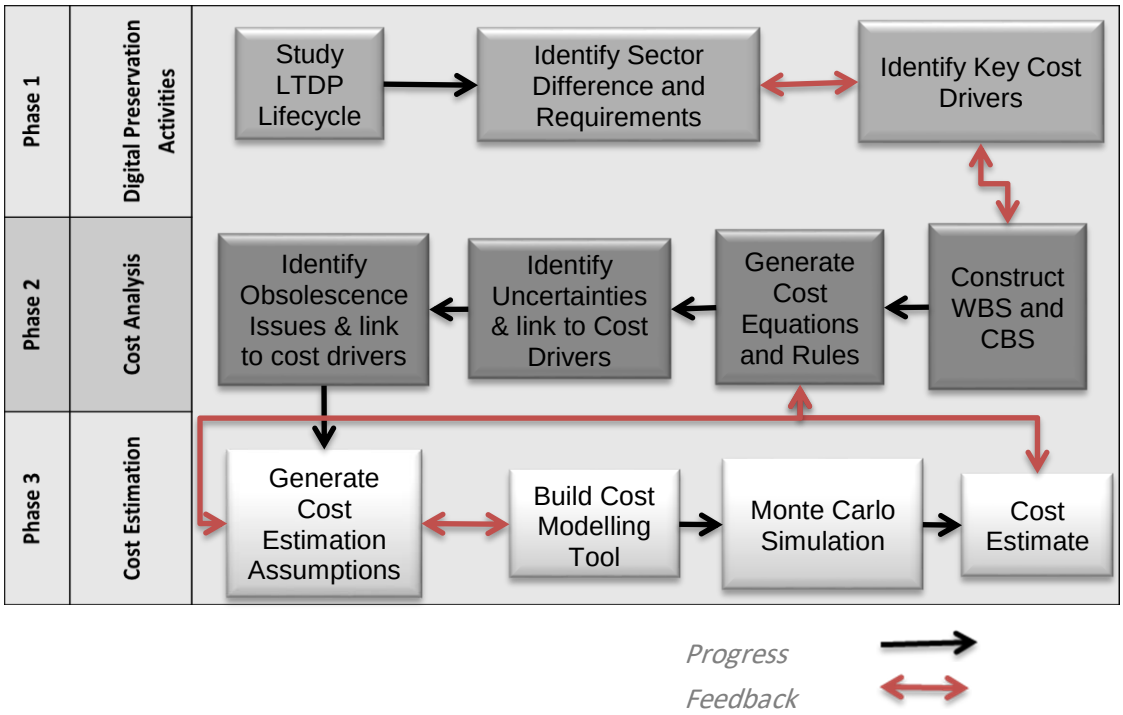


Figure 5-5 LTDP Cost Estimation Process

5.4 LTDP Cost Estimating Framework Construction

A framework explains main items to be investigated "*key factors, variables or constructs*" and the relationships between them (Miles, et al. 2014).
The developed framework, as shown in Figure 5-6, consists of three phases. The feedback points in the framework enable this design to automatically adjust itself over time to meet realistic conditions in an LTDP system. The first phase of the framework focused on the activities required by the user, the next phase generated equations and rules to enable the cost estimates in the final phase to be calculated.

Each phase of the framework was represented by several stages in the cost modelling process. There were 11 stages in total and each stage required some information from the user and/or from a previous process. Some stages contributed to a feedback system that may help a firm to continuously adjust its estimates to possible changes in its current or future LTDP plans. More details of the developed framework are given in the following sections.

5.4.1 Phase 1 – Digital Preservation Activities

As the first phase, Digital Preservation Activities is where the information about the future LTDP system is generated. A plan of requirements needs to be put in place and analysed according to company/firm policies and LTDP target.

First, the lifecycle of the preserved or "*preservable*" digital object should be thoroughly designed. This establishes a clear understanding of what activities should be carried out on the submitted and ingested digital objects. A well thought preservation plan will make:

- Better investment
- Well preserved digital assets
- Less unforeseen uncertainties or obsolescence issues
- Easier cost estimation
- Better cost control
- Easier future access to preserved information

This should be supported by a clear understanding of the business sector's requirements, which is matched with the business's own requirements, regulations and targets.

5.4.2 Phase 2 – Cost Analysis

Cost analysis is the second phase of the framework and within it the output of the resultant cost model can be modified, expanded and enhanced, from a single to a three-point estimate. A single point estimate is the calculated cost without considering any impact of obsolescence or uncertainty on cost and the result is a single number. A three-point estimate, however, provides the user with the minimum, the most likely and maximum estimated cost figures plus a probability distribution of these values. This is obtained by considering the impact of uncertainties and obsolescence issues on the single point estimate, resulting in a probability distribution diagram. Three-points estimates are so beneficial for company decision makers and policy developers which provides them with a probability for every case, best case or worst case. This will enable them to prepare financially and put technical checks that can reduce further impact on cost and of course on preservation quality.

5.4.3 Phase 3 – Cost Estimation

Within this last phase of the framework, the output is developed. Cost estimate is the outcome from this phase. By combining all the information generated from prior phases and adopting Monte Carlo simulation, a cost model can be developed that can provide a cost probability distribution and a three-points estimate.

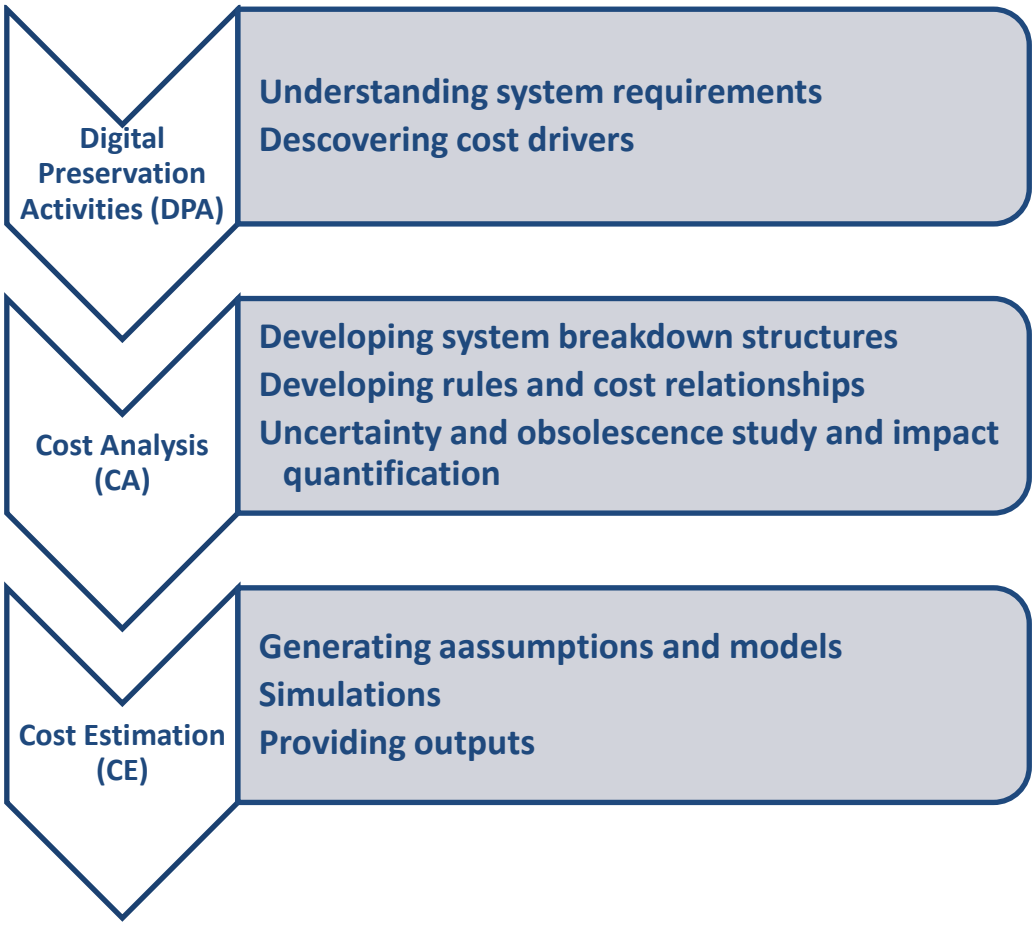


Figure 5-6 Conceptual LTDP Cost Estimation Framework

The framework is highly sensitive to preservation activities selected by the company/firm undertaking LTDP. Thus, showing the importance of a preservation

plan in the initial stages of LTDP system design for a specific organisation; where all cost information directly depends on this plan.

5.5 Chapter Summary

This chapter discussed exploring the findings of investigating possible uncertainties and obsolescence issues in LTDP systems in the chosen business sectors. All five categories of uncertainty discussed are common to many industries. The uniqueness of this research is that it shows that uncertainty categories exist in LTDP systems. Interviewed experts have agreed on the five uncertainties categories identified here, as shown in Figure **5-2**.

The scoring mechanism could not provide sufficiently accurate information, because the experts did not agree on a single value or a range of values, which are tightly packed. Hence, in both uncertainty and obsolescence scoring the Delphi method was used to reduce bias and subjectivity. It was clear from observing this exercise that most of the experts and interviewees who had joined the session were very pessimistic about the impact of cost. This shows two things: first, their worry that the funding of their work will not cover the mitigation of risk. Therefore, higher impacts were shown in most cases; and, second, their bias towards higher cost values due to their increasing concern over the sensitive information that most of them handle every day.

Adding human skills and preservation strategy obsolescence to the well-known issues of hardware and software obsolescence proved successful with the attendees. Most of them had not recognised that both should be dealt with, but

they unanimously agreed that they should be added to the obsolescence taxonomy.

In recognising the high impact of hardware and software obsolescence issues, one must take account of the high probability that both issues will arise during a preservation plan. These issues are usually mitigated successfully, but occasionally something goes wrong and an obsolescence issue meets its deadline. Unfortunately, recovering from this is generally very expensive.

The second half of this chapter discussed the cost estimation process and framework. Both were developed from the cost model that was initially generated as a single-point estimating model, then expanded to be three-points estimating model. The process is constructed of eleven processes, which enable the process user to:

1. Break down LTDP requirements
2. Generate cost information
3. Expand cost information
4. Develop a functional and accurate cost model

From the process, a framework was deduced that capture the conceptual essence of the cost estimation process. Three elements are encompassed within the framework:

1. Understanding Digital Preservation Activities required
2. Analyse information within the designed digital preservation activities to find cost sensitive information
3. Generate a functioning cost model

The next chapter outlines the validation process for the developed framework and extracted results. The chapter also validates a concept tool that was developed and tested especially for this research project. Developing this tool also meant developing some default values and generating a scenario generation to suit the relevant business sectors. The tool was tested by three teams, each from a business sector and was tested by the values expected from using a private cloud.

6 Validation of Long-Term Digital Preservation Cost Modelling Framework

6.1 Introduction

This chapter investigates the validity of the suggested LTDP cost estimation framework presented in Chapter 5.4. Case studies, shown further in this chapter in section 6.4.5, from the business sectors under investigation and results from different validation runs are presented in this chapter. Investigating the validity of this framework will help the reader understand the depth and novelty of the present research, whose limitations and line of probable future development are presented in the next chapter.

The first section of this chapter discusses the levels of validating the cost estimation framework through workshops with experts. This section includes validating the flow and single point estimation processes of the framework. The second section targets the validation of the uncertainty processes of the cost estimation framework. Following this, the third section explores the validation of the obsolescence process of the framework. These first three sections depend on the answers to the semi-structured questionnaires answered by experts, with a detailed discussion of each particular case and question. The results of these questionnaires along any comments received from the experts are also highlighted.

The fourth section of this chapter describes the concept proofing tool that was developed, which made possible the quantitative validation of the cost modelling framework. Each input, output, assumption and calculation of the tool is fully explored in this section.

The closing section to this chapter is a quick summary of what the chapter discussed. The summary reviews the outcomes of each section and revisits some of the important findings from validating the cost estimating framework.

6.2 Different Validation Tiers

The framework had to be validated by different experts in different set-ups and through different means of experiencing the outcome of this research. Different experts from the LTDP domain and different meeting set-ups helped to reduce positive or negative bias and may also have reduced the influence of opinionated experts who sought to direct the opinion of the majority in this closed validation session. To accommodate the experts, three different tiers of validation were designed that could pin-point the merits and otherwise in the design of this framework for estimating the cost of long-term digital preservation.

The first group consisted of highly skilled preservation experts from across Europe, along with experts from the three candidate business sectors; this was meant to ensure that the outputs and outcomes, of the framework would meet the requirements and expectations of the business sectors in question. This group of experts met weekly to discuss and improve the weekly progress of the design from the early-stages of constructing a framework outline to the final stages of discussing different uncertainties and obsolescence issues that should be represented in the output cost model. The validation meetings were crucial in adjusting the compass for the researchers so that they would not diverge from the core requirements of the targeted business sectors and in underwriting the compatibility of all results with the OAIS reference model (CCSDS, 2002).

The second group of experts were on a training course for LTDP practitioners. Three sessions were booked, and the framework and its outputs were discussed. Questionnaires targeting every process and sub-process in the design were checked to ensure that the design would lead the user to a stable and reliable cost model. The results collected from answers were freely discussed immediately afterwards and feedback was collected.

The third group of LTDP experts was interviewed individually by telephone. A presentation of the framework design, processes and outputs was sent to each of them by email, along with some questions. All the questions, discussions and results from these different tiers of experts were collected and compared.

Each group had important contributions, criticisms and positive feedback about the framework. After a fuller description of each tier, all their validation results are presented.

6.2.1 Weekly Validation meetings

From the beginning of the design phase for the framework, a continuous validate-update strategy was adopted (see Figure **6-1**). All the data collected contributed directly to the design and development of the LTDP cost estimating framework. Each new layer of design was put to the test by presenting them to a group of industry experts (see Table **6-1**), that met every week. This group Industry experts discussed the new design and gave their opinion on its validity of the design. Either some amendments to the submitted design were triggered or the next step in constructing the framework was taken.

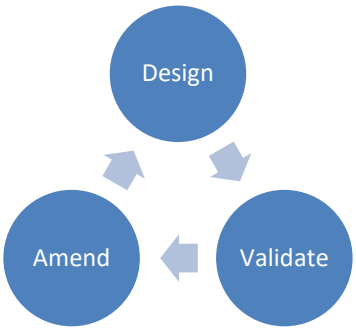


Figure 6-1 Validation: an integral part of the design

The experts' evaluation of each part of the framework, the process or the general flow of processes ensured a better cost estimate solution for the LTDP needs of the chosen business sectors. Giving representatives from these business sectors (see Table **6-1**) a crucial role in those weekly reviews made sure that the final outcome could theoretically be as close as possible to the targets.

Table 6-1 Experts Attending the Weekly Validation Meetings

NO.	COMPANY	POSITION	YEARS OF EXPERIENCE
1	IBM	Lead IT Storage and Systems Developer	24
2	IBM	Cloud Platforms	32
3	STFC	e-Science Developer	18
4	STFC	Library Systems	19
5	Tessella	Project Manager/Digital Archiving Solutions Development	22
6	Custodix	Digital Security Developer	11
7	CSISP	Pharmaceutical information Representative	3
8	Maccabi	Head of Healthcare Digital Technology	14
9	Maccabi	Head of Healthcare Digital Security	18
10	JRC	Head of Financial Digital Data R&D	19
11	Phillips	Digital Pathology - Developer	7
12	ATOS	Head of IT Systems and Configurations (ENSURE Project)	14
13	Frauenhofer	Healthcare Information Systems	23
14	Luleå University of Technology	Preservation Plans Suitability and Quality	5

The weekly meetings took the form of online conference calls at least an hour long. Experts were usually presented with updated developments to the framework and the resultant models for improvement. A series of validation questions suitable for each stage of the framework design were asked, to make sure that every step of the design was in line with their expectations. In the early stages, the improvements were usually major. The early design of the framework did not incorporate phases as in its final form, shown in Chapter 4 **Error! Reference source not found.**. Breaking down the framework into its elements was a task that confused the experts and many suggested that the processes in framework should be bundled in phases. This in turn helped reduce the difficulty of tackling the framework and in turn reduced the time required to reach a valid, usable and accurate cost model.

Another important validation input of the experts was that they made sure that every design step was consistent with the expectations and requirements of the targeted business sectors, especially in validating a unified LTDP whole lifecycle diagram, Figure **4-1**, that met their common demands and represented accurately their LTDP system designs.

Having experts from diverse backgrounds in digital preservation also generated conflict in validating some framework outputs, such as the breakdown of the differences in the LTDP requirements of the business sectors (see Figure **4-2**). The main reason that this forms part of the task of the framework is to highlight what the framework focuses on and to shed some light for users on what their business sector needs to know before continuing with outputs from the

framework. This would enable users to extract critical cost elements and key cost drivers, which is the key process in phase one of the framework.

When validating this process, 4 out of 14 experts requested the addition of a seventh requirement, that is “Company Rules and Regulations”. After many iterations, however, the seventh requirement was removed, since it diverts attention from the sector as a whole and focuses on the company/firm alone. Although the addition of a seventh requirement was rejected, all the experts agreed that it would not improve the framework’s outputs.

The final correction through validation was to separate the identification of obsolescence issues from identifying uncertainties in an LTDP system. The reasoning behind this was to highlight the importance of obsolescence issues and their impact on the output cost model. In combination with other things, it was very difficult to categorise uncertainties from obsolescence, since obsolescence is itself uncertain, but it has so important a role in digital preservation that its sub categories need to be clarified. This taxonomy of obsolescence not only helps to find more accurate outputs from the framework and in turn the cost model, but also highlights all the obsolescence issues that can impede the digital preservation plan.

Scrutinising phases, processes and flow in the framework was the first validation task, which was fortunately not only successful and satisfactory for all the experts, but also helped to make the framework’s more usable, accurate and user friendly. This group of experts contributed to validating all of the components of the framework, including impact of uncertainties and obsolescence issues on cost.

Some very positive feedback came from experts about framework design in the final week validation meeting. An IBM IT and Systems lead developer said that the work was *"very impressive*".

6.2.2 Validation via Industry Practitioners

Three 60-minute validation sessions were booked in the first Digital Preservation Advanced Practitioner Course by the Digital Preservation Coalition at the University of Glasgow. All the sessions were attended by 18 experienced practitioners in the LTDP field, who together had a spectrum of experience ranging from 1 year to 15 years (see Table **6-2**).

Table 6-2 Experts Attending 3 Validation sessions

No.	Company	Job role	Years of Experience
1	CJ-IMC	Independent Consultant	10
2	Tell Berlin	Research Data Coordinator	3
3	NLS	Digital Preservation officer	13
4	Digital repository of Ireland	Digital Archivist	1
5	DANS – Data Archiving and Networked Services	Data Manager	6
6	FRD	N/A	10
7	UC3M	Assistant Prof. Digital Libraries Masters Course	3
8	An Educational Institution	Librarian	1
9	Digital Preservation Coalition	Executive Director	15
10	CINECA	Software Integrator	2
11	Charles University in Prague	System Administrator	5
12	Digital Preservation Coalition	Senior Project Officer	7
13	Digital Preservation Coalition	Senior Project Officer	1
14	The British Library	Head of UK websites Preservation	15
15	Fuse – Institute Berlin	Research Developer	3
16	University of Leeds	Digital Content & Repositories Manager	13
17	VIAA	Digital Archivist	1
18	KEEP Solutions	Innovation Director	7

The main questions asked were on validating the construction of framework, appendix A.7, A.8 and A.9,cost drivers and the key cost drivers generated from the framework. The results from these sessions indicate that the essential components and outputs of the LTDP cost estimating framework satisfied most of their expectations. These experts validated all the framework components except obsolescence, since they had contributed to the obsolescence research outcomes.

The gravest criticism of the framework was that in its data collection in the design phase it depended heavily on experts' opinions and inputs. Subjectivity can mislead users of the framework into inputting pessimistic or optimistic values in the cost model, which in turn leads to a pessimistic or optimistic estimate. To overcome this, in developing the framework values, the researcher supplied the framework users with as many of the values as possible.

However, more positive feedback for the framework design and outputs was received. All the experts agreed that this framework was flexible enough to be adapted to many more business sectors.

6.2.3 Validation with Long-Term Digital Preservation Experts

Individual telephone interviews were carried out with 13 experts (see Table **6-3**). Each interview took at least an hour; some were as long as three hours. Experts were chosen to cover a wide range of understanding of LTDP systems, cost engineering and uncertainties in IT businesses. Once they had seen the details and capability of the framework, these experts gave very positive feedback in favour of the design.

Table 6-3 Telephone Interviews Validation Experts

No.	Company	Job role	I.T. Years of Experience
1	Honeywell	Electrical Engineer	5
2	Diversity NZ	Advisor/Investor/Commentator on Cloud Computing	18
3	WHO Geneva	Director Knowledge Management and Sharing	27
4	Accenture USA	Managing Director - infrastructure	21
5	Microsoft	Software developer	8 months
6	Stanford University	Digital Preservation Expert	12
7	IBM UK	IT Architect and Consultant	26
8	IBM UK	IBM Certified Client Executive	30
9	IBM UK	President, IBM Academy of Technology	29
10	STFC	SCAPE Project Mg	19
11	STFC	Project Mg in Data Center	17
12	NHS Scotland	NHS Scotland Head of Data Storage and IT Systems	30
13	Channel 4	Business Development Manager	10

The positive feedback from these experts revolved around how detailed the framework was and how much calibration is available to a user of the output cost model. Another positive comment was that with minimum effort a user from a different business sector could adapt the whole framework to suit his or her own sector.

A section of the framework that clearly raised the highest interest was the segmentation of uncertainties in LTDP systems and their identification process, (see Figure **5-3**), and the obsolescence taxonomy (see Figure **5-4**). The interest was generated, as they mentioned, by the lack of research needed for costing

uncertainties in LTDP. Condensing the three validation tiers, the results from all validation sessions are next discussed.

6.3 Validation Results

In this section, the combined results and all the feedback points from the experts are discussed. 45 experts in total were interviewed and the following points were compiled to show the percentage of expert approval.

6.3.1 Approval of framework construction

The following table, Table **6-4**, shows the percentage of approval that the experts expressed of the framework design and its components. Questions were asked about every finding of the research, the suitability of framework for the task and the adaptability of the framework to other business sectors. Some aspects scored as high as 100% approval while other aspects scored as low as 82%, but no lower.

Table 6-4 Approval Rating of Framework Construction

Framework Design Aspects	Approval Percentage
Framework phases	44/45 (97%)
Framework processes	44/45 (97%)
Process flow	44/45 (97%)
Whole LTDP lifecycle	45/45 (100%)
Differences in Sector LTDP requirements	43/45 (95%)
Key cost drivers (Data Volume, Retention period, Cloud Model and Processing rate)	40/45 (88%)
Work breakdown structures	39/45 (86%)
Cost breakdown structures	38/45 (84%)

LTDP system on private cloud cost equations	43/45 (95%)
LTDP system on public cloud cost equations	41/45 (91%)
Equations units and assumptions	45/45 (100%)
Categories of Uncertainties	42/45 (93%)
Uncertainty identification process	40/45 (88%)
Uncertainties in LTDP	41/45 (91%)
Obsolescence Taxonomy	44/45 (97%)
Obsolescence equations	37/45 (82%)
Uncertainties and obsolescence impact factors on cost	39/45 (86%)
Framework validity for task	45/45 (100%)
Suitability of framework for business sectors	45/45 (100%)
Adaptability of framework	40/45 (88%)
Framework respect of the OAIS reference Model	45/45 (100%)

Achieving over 80% in every design aspect of the framework reflects its solid construction and representation of real-world activities for LTDP systems. This is supported by the British Library's presenter on the digital practitioners' training course "***This cost model is one of the most detailed cost models for LTDP systems … similar to the LIFE model of the British Library***". In addition, the framework scored 100% approval for respecting the OAIS reference model and the expected whole LTDP lifecycle.

The lowest score in this regard, 82%, was for obsolescence equations, where some experts reflected their concern about treating obsolescence as a standalone requirement, rather than making it one of the technological uncertainties.

6.3.2 Experts' Comments and Feedback

A discussion was carried out after every validation interview/session, mainly to elicit the reasons for approving or rejecting parts of the framework. Their comments regarding construction and design of the framework are discussed next showing the rationale behind some of their choices.

The phases, process and flow of the framework were questioned by one expert from STFC, who commented that some processes could be combined to improve the flow of framework use; he gave the example of combining obsolescence with uncertainties, which, as mentioned, was meant to simplify the task of developing a general breakdown of uncertainties. Naturally, the same expert was a member of the group which recommended combining uncertainties and obsolescence issues.

All the interviewees accepted the need to preserve a file's entire lifecycle in a preservation system, obviously; all agreed that the framework components and mechanics were inspired and followed the OAIS reference model construction. One cost driver, however, the IT system processing rate, raised some concern from the experts; they debated whether it was a key cost driver or merely a cost element. The fact is that a faster processing rate impacts on the overall cost of any LTDP system, especially when backup exists or multiple fixity-check runs are required.

One of the biggest discussions concerned the WBS, CBS and the generated cost equations for private and public cloud models. The main point was security checks, which can be seen to recur in many parts of the breakdown structures and equations. Not only did the security repetitiveness raise vigorous debate between the experts, but different layers of security were also brought into discussion. The main reason was that the experts feared that repeating costs would be accumulated and would misleadingly skew the total cost. Accumulated redundant security activities are in fact isolated in the equations unless called for.

Uncertainties in the LTDP's categories and identification process raised a discussion about which exist in the public cloud deployment model and which in those of the private cloud and whether their impact factors will differ from one business sector to another. The obsolescence taxonomy was commended for incorporating preservation plan obsolescence, which led one expert to admit that he had first questioned and then approved it after discussion.

Finally, the experts showed their support for the suitability and validity of framework as a way of estimating costs for the LTDP system, and also in the designated business sectors. Most of them agreed and discussed the adaptability of the framework to other business sectors, though some questioned it, raising doubts about the usability of cloud computing in other business sectors.

6.4 Tool Long-Term Digital Preservation Cost Estimation Framework Proof of Concept

A tool was developed that can help to realise the concept behind the framework's mechanics and tests its outputs in real-world case studies. The main issue in designing a case study for this tool was the lack of historical data. Most

companies rely on a museum approach to preserving critical data and/or are not sure how to keep their data stable and accessible.

In one interview during the data collection phase of the present research, one manufacturer revealed that until the interview he had not realised the importance of LTDP. Certainly, most organisations now have more knowledge about handling digital data, but still, in the researcher's experience, more work on raising awareness is needed.

6.4.1 Design and Flow of the Estimating Tool

The tool is designed in Excel. An integral part of its conception is to be easily edited and adapted to needs of users. Figure **6-2** shows tool's main screen, where an inexperienced user can input main details required to produce an estimate.

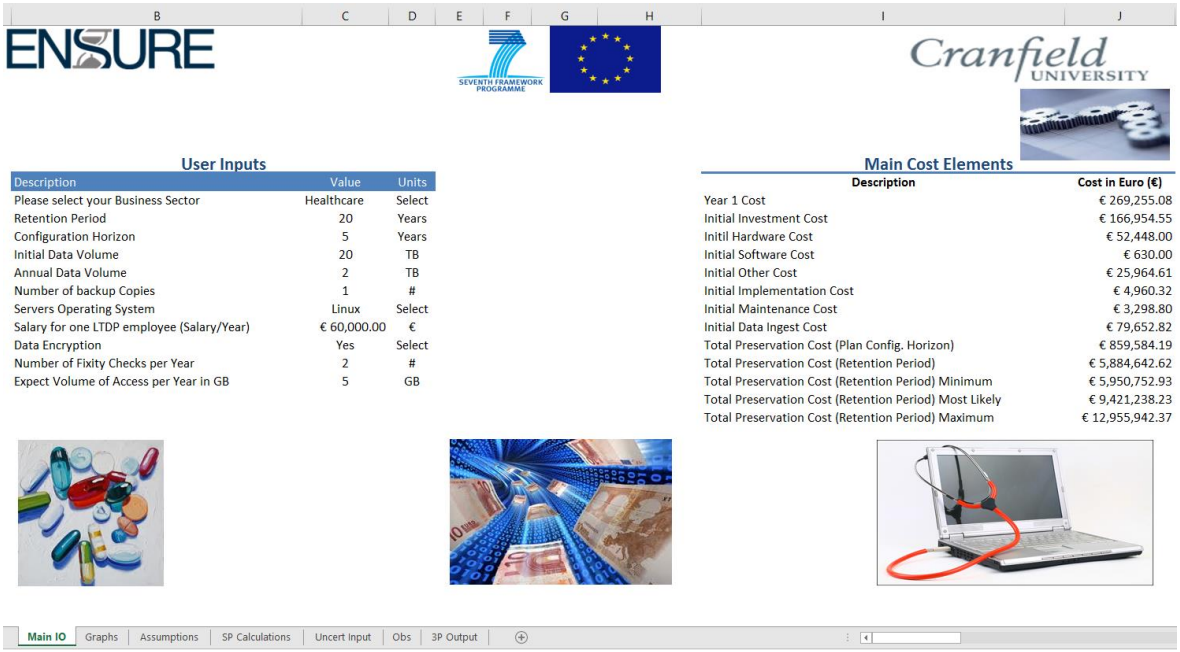


Figure 6-2 Proof of the Concept's Main Screen

The tool requires the user to input only 11 parameters unless any of the assumptions need to be changed. Those 11 inputs are:

1.Selection of the business sector 2.Retention period

3.Preservation configuration horizon 4.Initial data volume (Day 1 input volume)

5.Annual data volume 6.Number of required backups

7.Compute servers operating system 8.Salary for the LTDP employee

9.Whether encryption is required 10.Number of Fixity checks/year

11.Expected volume of accessed data/year

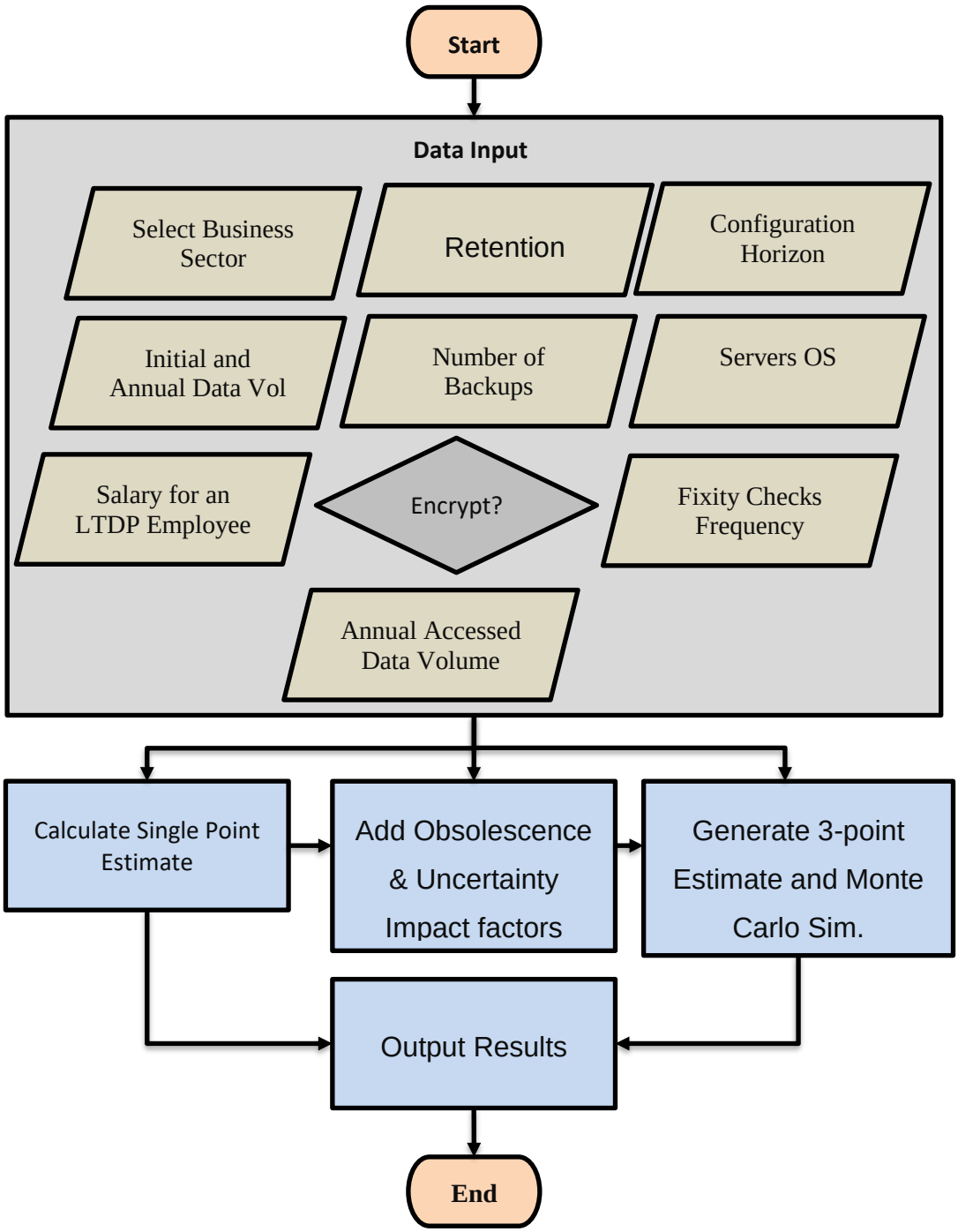


Figure 6-3 Tool Flow

The sector selection will impact on the complexity of the preserved files; for example, the data from healthcare firms are up to 3 times higher in complexity than any others, and the lowest in financial data. This is due to the use of many proprietary file formats and many video and picture formats. The configuration

horizon will help estimate the cost of the period over which the configuration plan is designed to work before being reviewed. Initial data volume is the volume of data that the organisation will load onto the LTDP system on its first day.

Choosing data encryption increases the processing time twice over, once in ingestion and once in accessing these data again. This is a measure of increased security. Finally, the data are expected to be accessed yearly; this is the whole purpose and the main cause for an LTDP system, activating most of the retrieval requirements and causing disruption. The design of flow of the tool is demonstrated in Figure **6-3**.

6.4.2 Assumptions

To generate a reliable estimate from an easy to use tool was challenging, since not all test users were ready to amend a lengthy spread-sheet. This led to making many editable assumptions, Figure **6-4**, that can be changed by a user who is willing to adapt the tool to his or her very specific case. These assumptions are by no means unreliable: they were designed under the supervision of representatives from all three of the chosen business sectors and are considered the basis for the case studies fed into the tool. These assumptions are as follows:

- Existing building
- Existing cables
- Compute Servers (CS) are IBM BladeCenter HS23E 8038B1G
- Storage Servers (Single Controller) (SC) are IBM System Storage DS3512 – Up to three expansion units and 12 drives

- Storage Servers (Expansion Units) (SEU) are IBM System Storage EXP3512 – up to 12 drives
- Rack Enclosures (RE) are IBM S2 42U Dynamic Standard Rack (99564RX) – up to 2 Chassis
- Chassis (CH) are BladeCenter E Chassis–86774TG– Up to 14 CS or 7 SC/SEU
- Hard disk drives (HDD) are IBM 2TB 3.5in 7.2K NL SAS HDD – 2 Tera Bytes
- Ethernet modules (EM) are 4x Layer 2/3 Copper Gb Ethernet Switch – one in each chassis
- Switches (S) are IBM RackSwitch G8052 (Front to Rear) – up to 48 devices
- Distribution switches (DS) are IBM RackSwitch G8124E (Front to Rear) – up to 24 devices, only if more than 1 switch.

Description	Value	Units	Value2	Units2
Numerical Assumptions				
Number of Working Days/Year	252	days/year		
Number of Working Hours/ Working Day	8	hrs/Day		
Rate of ingest/ 1 GB	9	Cplx	3	mins
Single Server Processing rate in GBs/Year working hours	13,440	GB/year		
Number of Hours per Year (24 x 365)	8,760	hours		
Single Chassis Power	2.32	kw		
Single compute server Power Consumption	0.1657	kW		
Single Storage Server Power Consumption	1.17	kW		
Switch Power Consumption	0.13	kW		
Distribution Switch Power Consumption	0.115	kW		
Working Day Cost	€ 396.83	€		
Other Calculations Assumptions				
Power Distribution System	€ 110.00	€		
25kVA Diesel Generator	€ 6,970.00	€	21	kW
75kVA Diesel Generator	€ 15,350.00	€	64	kW
120kVA Diesel Generator	€ 26,130.00	€	101	kW
UPS Price	€ 1,765.00	€		
5kW CRAC Unit Price	€ 3,010.00	€	5	kW
20kW CRAC Unit Price	€ 18,408.00	€	20	kW
42kW CRAC Unit Price	€ 23,343.00	€	42	kW
Security System Price	€ 1,640.00	€		
Fire Protection for 5X5 room	€ 14,000.00	€	25	m^2
Fire Protection for 10X10 room	€ 27,500.00	€	100	m^2
Fire Protection for 25X25 room	€ 55,000.00	€	225	m^2
Maintenance Assumptions				
Chassis Maintenance Cost/Server/Period	€ 1,970.00	€	5	Years
Compute Server Maintenance Cost/Server/Period	€ 531.00	€	5	Years
Storage Server Maintenance Cost/Server/Period	€ 3,114.00	€	5	Years
Distrib. Switch Maintenance Cost/Server/Period	€ 3,880.00	€	2	Years
Switch Maintenance Cost/Server/Period	€ 3,570.00	€	5	Years
Cooling Units Maintenance Cost/Server/Period	€ 2,602.00	€	5	Years

Main IO | Graphs | **Assumptions** | SP Calculations | Uncert Input | Obs | 3P Output

Description	Value	Units	Value2	Units2
Hardware Calculations Assumptions				
Total Data Volume	122,880	GB	20480	2048
Compute Server Price	€ 2,030.00	€		
Storage Single Controller Price	€ 3,350.00	€		
Storage Expansion Unit Price	€ 2,060.00	€		
Hard Disk Drive Capacity	2,000	GB		
Number of HDDs in Storage Signle Controller	12	#		
Chassis Price	€ 5,601.00	€		
Rack Enclosure Package Price	€ 6,820.00	€	0.72	m^2
HDD Price	€ 584.00	€		
Ethernet Module Price	€ 7,280.00	€		
Switch Price	€ 4,020.00	€		
Distribution Switch	€ 10,965.00	€		
Sofware Calculations Assumptions				
Windows Licence Price	€ 537.00	€	1	Years
Linux Licence Price / 3 years	€ 206.00	€	3	Years
Linux Licence Price / year	€ 68.67	€		
IBM Tivoli Storage Management	€ 218.00	€	1	Years
Cloud OS	€ 0.00	€		
Power Consumption				
Average Electricity Cost in Europe	€ 0.10	€/kWh		
Percentage Idle of Full load of IT systems	0.3	1=100%		
Cooling System Average Consumption/kW IT	€ 524.00	€/kW		
Idle Hours per year	8,760	hrs		
Processing Rates				
Ingest Collective Processing Rate (3GB/min)	60	GB/hr	180	Default Values
Fixity Check Processing Rate (1GB/min)	20	GB/hr	60	
Encryption/Decryption Processing Rate (1.5GB/min)	13.33333333	GB/hr	40	
Access Collective Processing Rate (3GB/min)	60	GB/hr	180	
Transformation Processing Rate (Assumed@ 3GB/min)	60	GB/hr	180	
Complexity of Clinical Trials	2.5			
Complexity of Healthcare	3			
Complexity of Financial	1			

Figure 6-4 Tool Assumptions

- UPS units (UPS) are IBM 2200VA LCD 2U Rack UPS (230V) (53952KX)
- Idle load of cloud system is 30% of full load.
- All monetary assumptions and working hours are in Euros and according to European average prices.
- Assumed 1 System admin to serve 280 servers or 140 storage nodes

All assumptions are integrated with the user's inputs filling the required elements in the equations.

6.4.3 Tool output parameters

The tool provides the user with 13 instant cost values, divided into two sections; initial costs and total costs.

- Initial Costs: denotes the day 1 commitment for the organisation, which includes the costs of year 1, initial investment, hardware, software, implementation, other, maintenance and data ingest.
- Total Cost: is the organisation's full monetary commitment, which can be used to develop a cost versus benefits model. This includes the total configuration horizon cost, which covers the period of the first preservation plan. It also shows the total preservation period cost as a single point and the minimum, maximum and most likely values for the total retention period (three-point).

A complete tab, Figure **6-5**, is another output of the tool that shows details of each cost element calculated annually and then aggregated together, forming cost values. This tab displays graphs of annual cost, configuration horizon annual and cost contribution of each cost element.

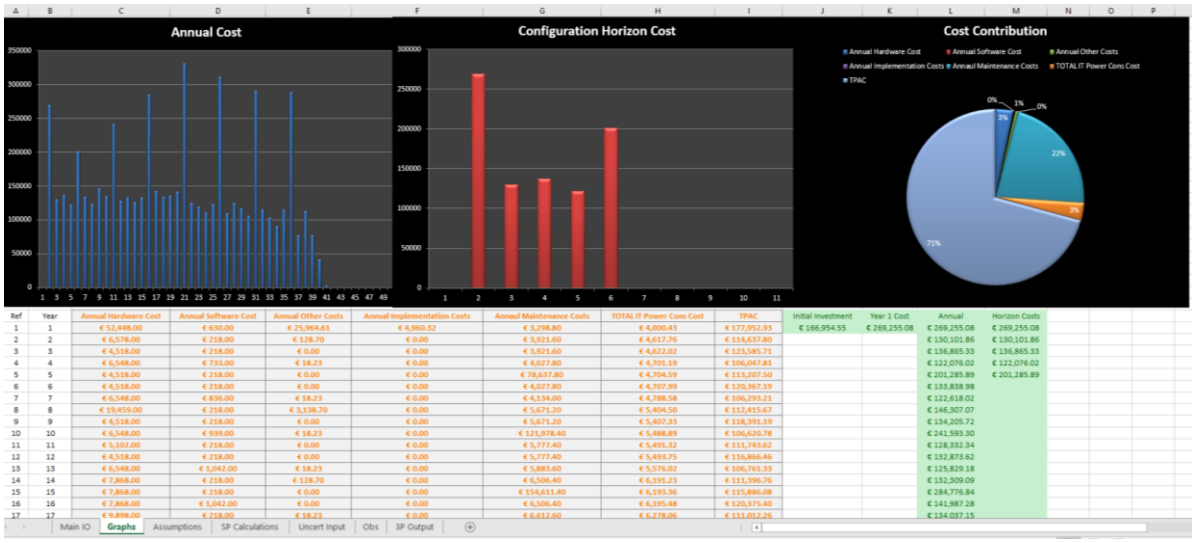


Figure 6-5 Output Costs and Graphs Tab

6.4.4 Tool Case-Studies

Table **6-5** shows the bases of the case studies used to test the validity of the tool and frameworks. With respect to all the assumptions mentioned above, the results of what was expected from the tool is based on the use by the three sectors of exactly similar hardware and software.

This is crucial since cloud systems' hardware and software can impact dramatically on performance and cost from a vendor to a vendor and from a model to a model. Since many compute servers consume variable amounts of electricity and give variable rates processing, this is something particularly important to keep in mind.

Table 6-5 Sector LTDP Requirements

	Healthcare	Clinical Trials	Financial
Preservation Duration	Forever to help with historical big data analysis. Effectively patient age + 25 years	15 years + any Promoter requirements	Client Data = Relation + 5 years. Market Data = 30 years
File Type	• Image • Video • Alphanumeric	• Image • Video • Alphanumeric	• Alphanumeric • Software
Access Rate	Very low; once or twice a year	Very low access, maybe none/year only on inspection. Inspection is 50 patients' data, varying from 0.5 – 1 GB.	3 yearly audits. 3 cases per audit.

The case studies were generated by experts from each business sector given the assumptions to work with, who were asked to calculate a single point estimate. Unfortunately, they could not provide historical data, since these were not available and their individual business sectors were unique. The input values used by experts to generate case studies are:

- A 20-year retention period with 5 years' configuration horizon
- 20 Tera Bytes of initial data volume and 2 Tera Bytes annually
- 1 backup copy with full encryption
- Linux OS
- Salary of €60,000
- Two fixity checks per year for whole data volume
- 5 Giga Bytes of accessed data annually

The values of the case study outcomes are shown later in the following section (tool validation).

6.4.5 Tool Output Compared to Case studies

Provided with their estimates, the tool's outputs are compared to case study values, Table 6-6, Table 6-7 and Table 6-8. The error percentage between the two values is recorded for all cost output. Of course, the values from each business sector will have different error percentages from the others, due to the differences between the sectors and the fact that the estimate is calculated by a different estimator. It is important to note that the cost values provided by the case studies were calculated without any knowledge of the framework equations or mechanics.

Table 6-6 Comparing Tool output to the Healthcare Case Study with Error

Output	Healthcare		% Error
	Tool	Case Study	
Year 1 Cost	€ 269,255	€ 237,000	12%
Initial Investment Cost	€ 166,954	€ 152,000	9%
Initial Hardware Cost	€ 52,448	€ 56,000	-7%
Initial Software Cost	€ 630	€ 570	10%
Initial Other Cost	€ 25,964	€ 24,000	7%
Initial Implementation Cost	€ 4,960	€ 4,400	11%
Initial Maintenance Cost	€ 3,298	€ 3,400	-2%
Initial Data Ingest Cost	€ 79,652	€ 75,000	5%
Total Preservation Cost (Plan Config. Horizon)	€ 859,584	€ 765,000	11%
Total Preservation Cost (Retention Period)	€ 5,884,642	€ 5,000,000	14%
Total Preservation Cost (Retention Period) Min	€ 5,950,752	€ 4,700,000	21%
Total Preservation Cost (Retention Period) Most Likely	€ 9,421,238	€ 8,300,000	12%
Total Preservation Cost (Retention Period) Max	€ 12,955,942	€ 14,500,000	-12%

In the comparison of Healthcare results, Table **6-6**, total preservation cost as a single point has an error percentage of 14% overestimation over the case study. From results, the error percentage averages at 8 % overestimate by the tool.

Estimates vary between an overestimate of 21% and an underestimate of 12%. The initial cost estimates hover around the error margin of 10% overestimate, while the total cost estimates hover around the 17% overestimate.

Running a Monte Carlo simulation, Figure **6-6**, shows that a certainty level of over 80% lies between a minimum of €7,923,609 and a maximum of €10,971,832.

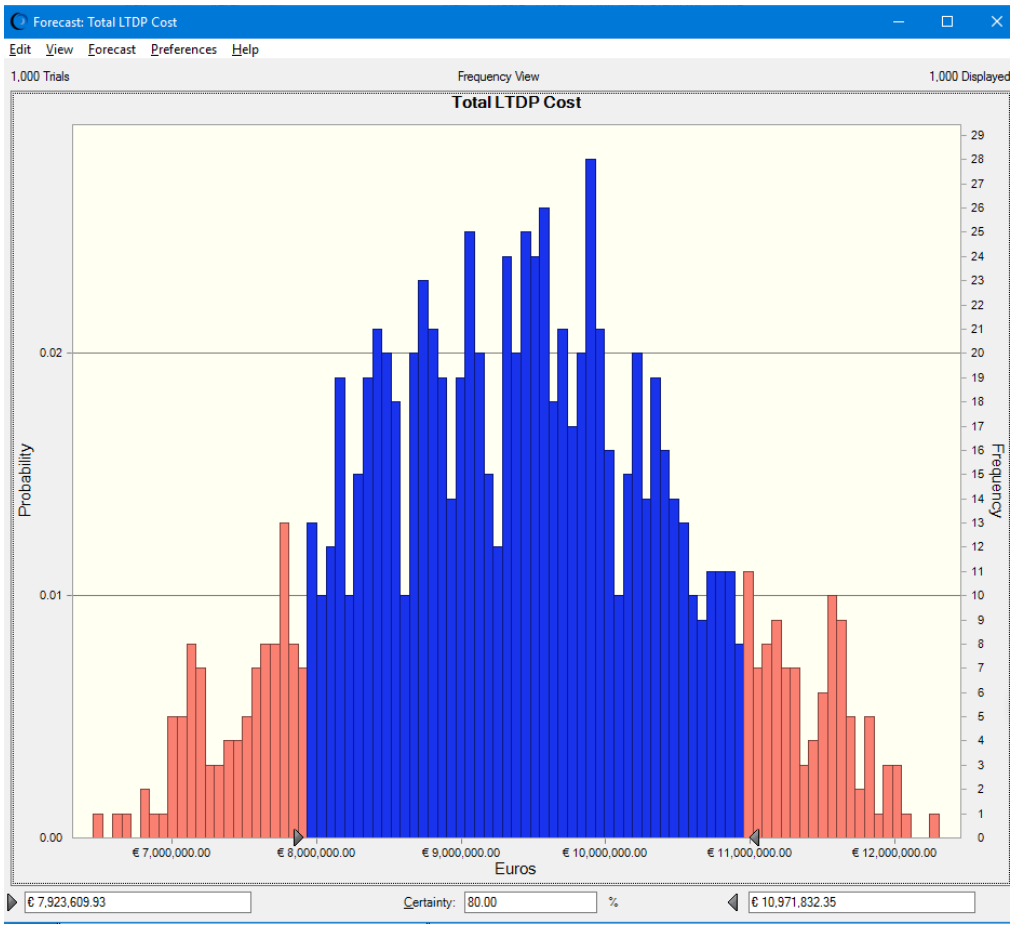


Figure 6-6 Monte Carlo Simulation for the Healthcare Total LTDP Cost

In the Clinical Trials results comparison, Table **6-7**, total preservation cost as a single point has an error percentage of 11% overestimation over the case study. From the results, the error percentage averages at 13 % overestimation by the tool.

Estimates vary between an overestimate of 25% and an overestimate of 3%. The initial cost estimates hover around the error margin of 15% overestimation, while the total cost estimates hover around a 20% overestimation.

Running a Monte Carlo simulation, Figure **6-7**, shows that a certainty level of over 80% lies between a minimum of €7,648,648 and a maximum of €10,575,924.

Table 6-7 Comparing Tool output to Clinical Trials Case Study with Error

Output	Clinical Trials		% Error
	Tool	Case Study	
Year 1 Cost	€ 286,552	€ 260,000	10%
Initial Investment Cost	€ 173,518	€ 166,000	5%
Initial Hardware Cost	€ 50,418	€ 44,000	13%
Initial Software Cost	€ 527	€ 450	15%
Initial Other Cost	€ 25,946	€ 22,000	16%
Initial Implementation Cost	€ 4,960	€ 4,000	20%
Initial Maintenance Cost	€ 3,192	€ 2,800	13%
Initial Data Ingest Cost	€ 88,473	€ 86,000	3%
Total Preservation Cost (Plan Config. Horizon)	€ 843,392	€ 800,000	6%
Total Preservation Cost (Retention Period)	€ 5,680,596	€ 5,100,000	11%
Total Preservation Cost (Retention Period) Min	€ 5,738,080	€ 4,400,000	24%
Total Preservation Cost (Retention Period) Most Likely	€ 9,083,411	€ 7,350,000	20%
Total Preservation Cost (Retention Period) Max	€ 12,490,831	€ 9,400,000.00	25%

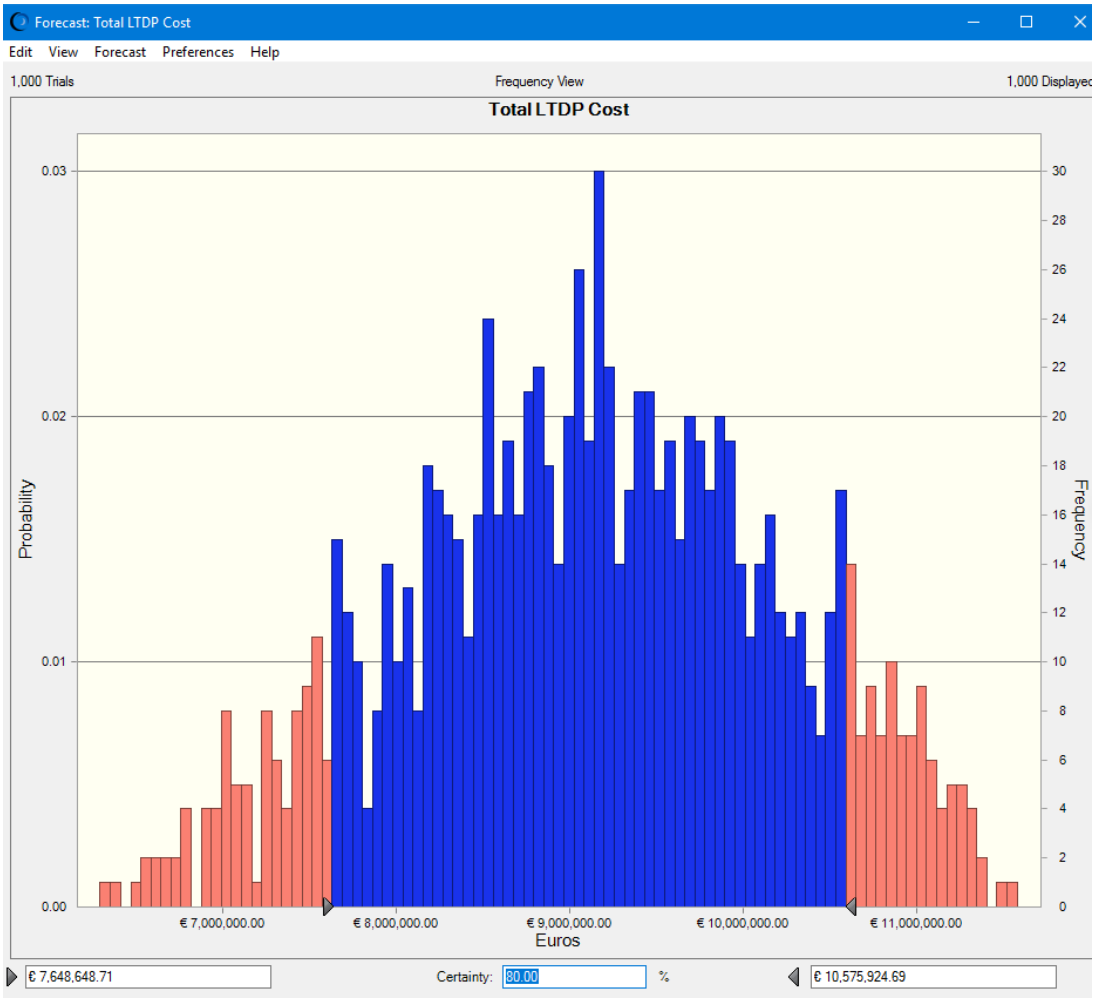


Figure 6-7 Monte Carlo Simulation for Clinical Trials Total LTDP Cost

In the Financial results comparison, Table **6-8**, total preservation cost as a single point has an error percentage of 6% overestimation over the case study. From the results, the error percentage averages at 16 % overestimate by the tool.

Estimates vary between an overestimate of 18% and an underestimate of 8%. The initial cost estimates hover around the error margin of 12% overestimation, while total cost estimates hover around a 9% overestimation.

Running a Monte Carlo simulation, Figure **6-8**, shows that a certainty level of over 80% lies between a minimum of €6,736,648 and a maximum of €9,177,997.

Table 6-8 Comparing Tool out to Financial Case Study with Error

Output	Financial		% Error
	Tool	Case Study	
Year 1 Cost	€ 204,951	€ 198,000	4%
Initial Investment Cost	€ 135,853	€ 123,000	10%
Initial Hardware Cost	€ 48,388	€ 40,000	18%
Initial Software Cost	€ 424	€ 360	16%
Initial Other Cost	€ 25,928	€ 25,400	3%
Initial Implementation Cost	€ 4,960	€ 4,400	12%
Initial Maintenance Cost	€ 3,086	€ 2,600	16%
Initial Data Ingest Cost	€ 53,066	€ 52,000	3%
Total Preservation Cost (Plan Config. Horizon)	€ 667,879	€ 620,000	8%
Total Preservation Cost (Retention Period)	€ 4,982,928	€ 4,700,000	6%
Total Preservation Cost (Retention Period) Min	€ 5,000,780	€ 4,350,000	14%
Total Preservation Cost (Retention Period) Most Likely	€ 7,912,082	€ 7,200,000	9%
Total Preservation Cost (Retention Period) Max	€ 10,885,386	€ 11,700,000	- 8%

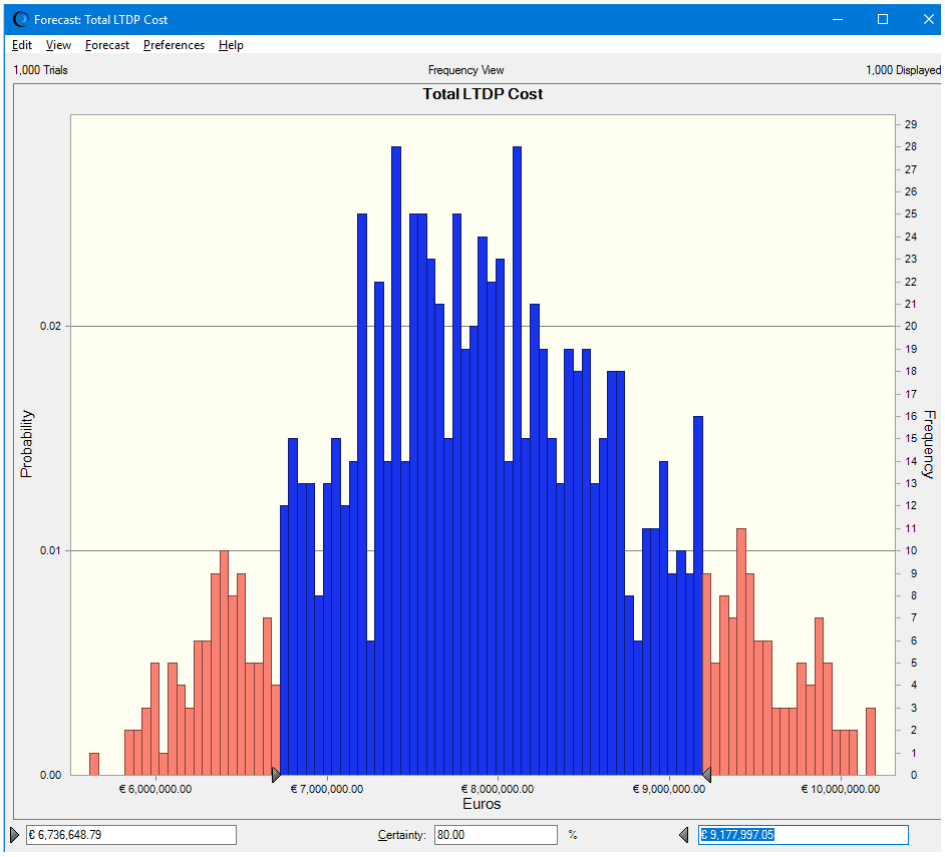


Figure 6-8 Monte Carlo Simulation for Financial Total LTDP Cost

6.5 Summary

The validation process of the framework design was split into three main phases, primarily for continuous improvement, by monitoring the quality of the research weekly. After confirming a solid foundation, a series of group and individual interviews with semi-structured questionnaires was carried out, targeting the rigorous testing of the theory behind the design of the framework. Many issues were detected through the weekly validation meetings; hence in the final validation interviews the conflicting issues were minor and with a discussion session after each interview, most of the negative feedback was withdrawn.

Following the validation interviews came the final stage of validating the framework's feasibility. Designing an Excel based tool to calculate the costs for an LTDP system, based on the private cloud deployment model, demonstrated that the framework could deliver results very close to real-world results. With an average error percentage of 10%, industry experts deemed the design feasible for use and representative of business sector requirements to the utmost detail.

The European reviewer who accepted the cost model design said "*I want to congratulate all the members for a successful project. I was actually impressed with the technical output of the project. The technical part was fantastic. I want to congratulate you for the way you have worked together*".

7 DISCUSSION AND CONCLUSIONS

7.1 Introduction

This chapter presents an overall discussion and draws some conclusions from the whole research project. It discusses the stages of developing a cost modelling framework to serve selected business sectors, estimating the costs entailed by long-term digital preservation. The chapter begins by discussing the research findings (see section 7.2) which focuses on the literature review, research methodology and research results. Then section 7.3 looks at the research and key outcomes that define the main successes in undertaking this project. The contribution to knowledge of the present research is covered in section 7.4, leading to a discussion on fulfilling the aim and objectives of the research. Finally, section 7.6 concludes both the chapter and the thesis, apart from a look at the study's limitations and some suggestions for future work in section 7.7.

7.2 Discussion of the Research Findings

7.2.1 Literature Review

Chapter 2 covered a survey of the literature in three research domains, namely, digital preservation, cost estimation and cloud computing. These are the main research areas requiring more information before the unknowns of this project could be explored. Digital preservation, to start with, requires some understanding of the different driving forces that pushed the scientific community towards investing in long-term digital preservation. Investing in long-term digital preservation systems and technologies depends on the value of the preserved information. Since the dawn of human civilisation, our species has thrived and progressed only on the basis of well-kept and understandable information. Over

time, much information has been destroyed, damaged or never retrieved from records that over time had become indecipherable, causing the loss of many useful items of knowledge.

Now in the digital era, some experts consider digital records to be fragile by nature. One author has gone on to compare this digital age with the "*Dark Ages*". The main points about the inherent fragility of digital data cannot be gainsaid, but the potential in maintaining a very high percentage of generated information is huge. Not only this fact, but also the chance to be extremely selective in what is preserved encourages entities to keep their most useful and valuable data in a preservation-friendly manner.

Enriched security measures, connectivity speeds, standardisation and the ease of creating multiple copies are points that seem to set this digital age as a cornerstone for future technological triumphs, while acknowledging the fragility of all kept information.

Four preservation techniques were identified in the literature: the museum approach, emulation, encapsulation and migration. The museum approach keeps the old technologies to retain as much as possible of the original "feel" of the information, with the risk of losing total access when this old technology collapses, due either to technical failure or the loss of the user's ability to use the technology. Emulation, as its name suggests, allows the preservation system to imitate the previous environment of the data within the new one. This can still provide the user with an experience that is very close to the original "feel" of the preserved information. In encapsulation, the preserved data are encapsulated

with as much information as possible to help people to access it at some future time, while in migration the data and its dependant technologies are both regularly updated to newer and stable versions, which may impact on the original “feel” of the data over multiple cycles of migration.

Any functionally reliable long-term preservation system usually relies on more than one preservation technique. Selecting those techniques follows some rules and standards, put together to ensure that most users have a reference to return to. A very important standard used by most preservationists is the Open Archival Information System reference model. It sets a path for the user, from the first step of preparing a set of information to be preserved, through preserving and archiving the information and maintaining a healthy archive to finally retrieving quality, accessible and understandable information from what was preserved. The OAIS also suggests some administrative protocols and some mechanisms covering access rights. Other standards apply to certain sections in architecture of a preservation system: PREMIS (Riley, 2007), for example, regulates metadata implementation strategies; NESOTR sets a best practice guide for managing digital storage (DCC 2010a); DRAMBORA has been devised for digital repository auditing (DCC 2010b) and TRAC certifies a repository or a digital facility (DCC 2010c).

Estimating the costs for long-term digital preservation systems has accompanied the research on preservation since its birth as an idea, especially now that so many businesses are interested in preserving information. Commercial entities are always careful in predicting financial commitments; hence a number of projects have started to target the realisation of costs for long-term digital

preservation systems. Unfortunately for the commercial sector, the main beneficiaries from the number of projects started in late 2000s were the heritage, scientific community and library sectors, where governments and authorities had to understand how much money from taxpayers had to be invested. Hence, a research gap was found between the existing cost models and the requirements of the commercial sector.

The present heritage and library long-term digital preservation cost models, some more than others, mostly rely on or refer to some of the basic OAIS reference models; LIFE by the British Library is a three-stage project that focused on quantifying the cost of digital preservation for the library by adopting a lifecycle approach. CMDP deals with data from the Danish Royal Library and a consortium of universities and state bodies from Denmark. Both focus on the heritage and library sectors and use the ABC costing technique.

The science sector and universities found themselves in a similar position to the heritage sector, in holding or generating an enormous amount of very valuable information without fully understanding the cost of preserving it. NASA stepped in to solve the problem with CET, a very expensive toolkit developed over time by comparing 29 projects. It has a 22.9% error rating (NASA, 2011). Keep Research Data Safe (KRDS) was developed by Charles Beargrie (2010) and funded by JISC in the UK; this also depends on OAIS and targets the research data held in universities, hence its name.

This thesis takes another look at costing for digital preservation but this time for the business sector, while integrating cloud computing technologies and

considering the impact of uncertainties and obsolescence issues. Here all the components of this thesis are unique and innovative and should bring a universal answer to any sector wanting to preserve data. The innovation in this case comes from overcoming new and uncharted challenges, one of which is the preservation of healthcare data for commercial use for more than 100 years, alongside new technologies and sectors whose their business is not in any way related to archiving data.

Before this research project, targeted business sectors had no dedicated cost model or cost modelling framework to serve their needs. No cost model took into consideration modern technologies such as cloud computing in its different deployment models or any model that considered any skewing of its output, due to the impact of uncertainties and obsolescence issues. To plug these gaps, a framework was a suitable solution that would serve many business firms and would accommodate new technologies and impact factors in cost metrics.

Cloud computing was developed on the basis of the development of grid computing in the 1990s, where a network of computing nodes provides constantly available on-demand resources. Four available cloud deployment models provide users with a range of forms of availability to suit their business needs. Private, public, community and hybrid computing clouds are the ways in which cloud infrastructure is deployed. Private cloud ensures the geolocation, improved security and custody of digital assets for users, while public cloud is available to most people. Users can rent only the resources they need and do so on a pay-as-you-go basis; they find that the costs are lower but that security risks can arise.

Community clouds may be deployed, for example, between an exclusive number of users, sharing similar interests and the resources of the cloud Finally, hybrid clouds are composed of more than one deployment model; while they remain unique they connected through technology portals to ensure the free movement of information. Estimating costs for cloud computing relies on five important cost drivers: computing power and storage facilities, power and ventilation, the use of computing power, electrical consumption and traffic and network handling.

Modelling cost means predicting the cost of a set of tasks or activities before they occur. Cost modelling techniques have been set and debated by many authors and industry experts, but all agree that modelling and estimating cost and expenditure should contribute to improving performance and the handling of spending for companies that are interested.

Cost estimation techniques are generally classified as quantitative or qualitative. Quantitative methods analyse numbers, figures, the designs of products, etc.; whereas qualitative methods rely on comparing opinions, old and new products and experiences. From these two techniques four main sub classifications are identified: parametric and analytical under ‘quantitative methods’ and intuitive and analogical as ‘qualitative methods’.

From the analytical quantitative cost estimation techniques, activity based costing (ABC) was chosen for this project for several reasons. A lack of historical data significantly skewed this choice towards ABC; this system also allowed the tasks and activities to be broken down with a high level of certainty and it was also the

case that ABC does not seem like a black box to the inexperienced user. Moreover, ABC can be easily adapted and amended by future users.

7.2.2 Research Methodology

In Chapter 3 a detailed research methodology plan was set out. The project was designed around a qualitative methodology. Being qualitative has, however, the drawback that some of the information and data collected through researching could be vulnerable to positive or negative biasing from the interviewed experts. The researcher mitigated this uncertainty by increasing the number of experts interviewed and making sure that they came from different background within the research domain; and by applying the Delphi technique in some interviews where experts discussed their answer to reach an agreeable solution and by asking experts when possible to cross-check each other's information, so as to settle any discrepancies.

Interviewing for this project was done through a semi-structured questionnaire applied in:

- Face-to-face interviews
- Teleconference interviews
- Group interviews

Answers were validated through regular weekly meeting with experts and three case studies, one from each sector.

7.2.3 Developing a Framework for Estimating the Cost of a Complete Long-term Lifecycle of Digital Preservation

At its core, the framework is designed to predict the complete lifecycle cost of long-term digital preservation on cloud computing, including uncertainties and obsolescence. In Chapter 4 the processes and results of constructing a framework are shown, up to the single point cost calculation. The framework comprises three phases and 11 processes/stages in total. Phase one is to capture the digital preservation activities. It starts by studying the lifecycle of a file within a preservation system, followed by identifying the sector requirements and finally finding the cost drivers. The first phase outputs a detailed set of cost drivers and the expectations of the preservation system in the business sector. This provides very important data that are later used to construct a cost model: the years of preservation (retention period), volume of preserved date, rate of submission to the preservation system and cloud computing deployment model. Adding to these the expectations of the preservation system in the business leads to phase two of the framework.

The second phase digests the information collected in phase one and its stages and combines it all to generate work and cost breakdown structures and in turn cost equations. The work breakdown structure is found achieved by combining the digital preservation lifecycle expected, such as that shown in Figure 4-1, with the business sector requirements. Combining these results in the tasks underlying each part of the lifecycle generates a work breakdown structure. Looking further inside these work breakdown structures and seeing them in

conjunction with the cost drivers also generated in phase one produces a cost breakdown structure.

With a cost breakdown structure that represents the tasks and actions required, it becomes straightforward stage to generate the equations that correspond to each part of the breakdown. Using these equations by inserting actual real values and cost assumptions results in a single point estimate. Phase two targets an analysis of the cost of long-term digital preservation, and within it the study of uncertainties and obsolescence also is triggered in two stages, which are discussed in the next subsection, 7.2.4.

However, the single-point cost estimate that was produced can be used as a starting figure. From Chapter 4 the close relationship can be seen between the OAIS reference model and the produced work and cost breakdown structures. This is in line with standardising the expected systems with a tested and industry-leading model. Firms with regulatory bodies that monitor their performance according to industry standards find this very useful, thus avoiding any legal issues that may arise with digital preservation and keeping long-term custody of information.

In the second phase of the framework, it is clear that the cloud deployment model impacts greatly on cost equations; this is due to the nature of each cloud model. In this research project, the experts agreed only that two deployment models can be used within the business sector, namely, private and public clouds, depending on the sensitivity of data. So, as Chapter 4 shows, two different sets of equations had to be developed, using two different deployment models on a foundation of

the work and cost breakdown structures. This posed a challenge, since both sets of equations had to be rationally verified with experts who defended both deployment models.

In the third phase of the framework, some cost assumptions – for example, the cost a man-hour, the currency, basic computing component prices, etc. – had to be added to these equations to provide the user with a numerical output.

7.2.4 Quantifying uncertainties and obsolescence issues

Chapter 5 clarifies how uncertainties and obsolescence issues impact on the single point estimate that has emerged. An uncertainty identification process was developed to identify the uncertainties in long-term digital preservation systems. The user follows the identification process along with the generated work breakdown structure, to clearly understand what might impact on the work being done and start to formulate some contingency plans. Five categories of uncertainty can exist within a preservation system: economic, technological, business-related, physical and regulatory uncertainties. Each can occur on its own or with others and each has its own unique impact factor and probability of occurrence. Some uncertainties turn out to be risks that a user must mitigate and others are opportunities that the user should exploit for his own benefit.

As an uncertainty in itself, obsolescence was clearer as a concept to discuss with experts; as one of them commented, “The entire preservation system is built just because of obsolescence”. A detailed taxonomy was also developed to help users pin-point areas of cost generation. Mitigation strategies are always in place and those will add the costs that they incur. Interestingly enough, the

obsolescence of preservation plans themselves, though identified was not mentioned by most of the experts in the first run of questionnaires. After its introduction at a meeting, they mostly agreed that it existed and that effort had to be made to update the original plan if it became un suitable later on.

Combining those impacts with the single point estimate found earlier and in the final phase of the framework making a run of triangular Monte Carlo simulations, results in estimates of the maximum, most likely and minimum cost. his three-point estimate is the ultimate target of the present research, which combines within its values every stage and phase of the framework.

7.2.5 Framework Validation

In Chapter 6, a detailed validation of the framework was shown. The validation was done continuously while designing the stages and phases of the framework. Three tiers of validation meetings were held, to ensure that enough data were collected and that its analysis was in line with all the experts' understanding and views: a weekly validation meeting, validation via industry practitioners and validations with preservation experts. Each tier contributed to ensuring that the data and information collected stayed within expectations and that anything new that was discovered was neither random or unsupported.

The experts' approval of the framework components ranged from 82% to 100%; a high percentage can be reassuring especially in the final stages since the experts who contributed to the early stages of the research were not included when it came to checking for discrepancies and faults. What encouraged the experts to review the framework with such high percentages of approval was the

solid ties with the OAIS model. The OAIS model is considered a cornerstone in preservation systems, even with the concerns expressed about it at present.

To validate the framework against industry case studies, a concept proofing tool was developed using the framework's logic, phases and stages. The tool requested 11 items of information on the landing page, corresponding to the data required by the corresponding business sector and to the assumptions that were made, such as ignoring the cost of building rooms for a data centre or wiring rooms. Data centre machines were selected to represent the lowest tier acceptable in the industry; the estimate for them would change when different equipment was chosen for the data centre.

Most of the behavioural assumptions were made about the data centre staff followed the industry's best practices and advice from experts.

On average, the tool showed a 10% error from the case study information, which led industry experts to approve the framework; the design also received commendation from the European reviewer along with the commendation of a prominent industry expert.

7.3 Contributions to Knowledge

The main contribution of this research project may be summarised as follows:

- Cost Modelling Framework: Collating all these steps would lead the user to calculating a cost estimate that would be useful for their establishment, for which a framework was designed. The reason for making a framework the target from the beginning was that it made the main outcome from this research project very useful for more firms than had been specified and easily

adapted by them; that it increased the ease of personalisation per business sector and even per firm, company or organisation; and that it made it feasible to upgrade and adapt cost estimates to new challenges and technologies. The innovation within each phase of designing the framework is apparent, since it answers many questions for businesses that have been obliged to invest in preservation systems. It presents the preservation community with a framework that is malleable enough to be adapted by many firms, but still solid in its output resolution and accuracy. A concept proofing tool was also used to validate the entire output of the framework and to provide a deeper, more engaging experience from using the framework step by step.Long-term digital preservation Uncertainty Identification Process: Understanding the nature of uncertainties leads to a deeper understanding of ways to detect them and assess their direct and indirect impact on cost. As one of the uncertainties, obsolescence itself was also thoroughly studied and detecting it led to a taxonomy of the different obsolescence issues found to impact on long-term digital preservation systems and the accurately calculation of its direct impact on a single point estimate. Integrating the impacts on a single cost estimate from uncertainties and obsolescence issues generates a three-point estimate.

- Long-term digital preservation Obsolescence Taxonomy: A detailed taxonomy of the identified obsolescence taxonomy was developed, to help users generate mitigation data, which by turn would affect the future costs of obsolescence and the way in which they would be estimated.
- Impact of Uncertainties and Obsolescence on the Costing of Long-Term Digital Preservation Systems: when all the issues and their direct impact on

cost drivers had been collected, and measured, the three-point estimate was calculated by implementing this impact in the framework. This flexibility in the final framework design meant that all this information could be re-measured and updated in the future.

- Single Point Cost Estimate for long-term digital preservation systems, from a full single point cost model that represents an actual and realistic long-term digital preservation system design especially built for the three specified business sectors to understanding what uncertainties exist in such systems and what the impacts of this single point estimate are.
- Three Point Cost Estimate for Long-Term Digital Preservation Systems: by combining the single-point estimate and the impact factors of obsolescence and uncertainties, framework users could put together a three-point cost estimate. This is especially important to users with a commercial background coming from the business sector.

All these solid research results were accompanied by a three-stage validation procedure and a concept proofing tool that was also tested for ease of use and intuitiveness. Any industry expert who joined any design phase was automatically excluded from validating it. All the experts were asked to answer at least two semi-structured questionnaires about validating single-point estimates, uncertainties and obsolescence, three point estimates and the full framework.

7.4 Fulfilment of the Research Aim and Objectives

This research project aimed at seven objectives. These, mentioned in Chapter (1) section 1.4, were to find the total lifecycle cost of a long-term digital

preservation system. All these objectives have been secured and in this section the achievement of each is discussed in turn.

To understand practices and processes of long-term digital preservation, the literature was intensively reviewed (see Chapter (2), where all the aspects of the research topic were studied). Cost estimation, digital preservation and cloud computing were pursued. The literature was in reputable papers and books on one side and the findings from it were supplemented by a series of interviews with industry experts. The experts 'contribution cleared up many of the ambiguous areas found in the various textual sources. Other research topics, such as uncertainties and obsolescence, were further explored by attending conferences which helped to uncover some useful techniques for handling both topics in the context of digital preservation.

Understanding the digital preservation domain revealed companies' requirements from a preservation system. Business sector representatives were very clear on what their expectations were from such a system or service. Even though they could not satisfy their preservation targets, they could pin-point their needs, making it straightforward to construct a list of their requirements. Normalising it with the industry standard took some research, but was achieved.

A work breakdown structure is needed to reach a cost breakdown structure, and both were realised by combining the knowledge gained from the literature review with the industry requirements and standards, especially the OAIS standard. Breaking down tasks made it easier to reach the next objective, of identifying a cost estimation technique suitable for this project. Tasks were easily broken down

into activities, resulting in an activity based costing; this ensured that the results from the framework were also easily manipulated by users from the targeted business sectors, who are not essentially oriented towards cost estimation.

The uncertainty identification process was designated a stage within the designed framework. The process functions in the framework as an integral process of quantifying the uncertainties and integrating the result with a costing for the whole system. Part of the innovation in the process is its capacity to differentiate between risks and opportunities; it directs users to mitigating or exploiting them accordingly; which leads directly to understanding obsolescence and generating its taxonomy.

Using data from uncertainties identification process and obsolescence issues taxonomy enabled their impact on cost drivers to be calculated and incorporated, as shown in Chapter (5). Interviews with industry experts pointed out the literature needed for quantifying uncertainties and obsolescence techniques, thus reaching a combined solution for assessing their impact within the design of the framework. This was done while considering the ease of use and amendment by users.

All the results and goals of the previous objectives were combined in a framework design which delivered the validated results, satisfying the business users from all the targeted business sectors, as demonstrated in Chapter (4).

7.4.1 Research Outcomes

In this section of the chapter, main research outcomes are discussed and presented in accordance with the structure of the thesis. The discussion adds a consolidating element to the results of this research project.

Six outcomes resulted from carrying out this research project, establishing a satisfying answer to the research questions and attaining the aim of the research. First, targeted business sectors' long-term digital preservation system requirements were identified, which supported any design of the system with the knowledge of how it is expected to perform, with what inputs/output and for how long.

A digitally preserved file lifecycle was a basic input to elicit the key cost drivers and the breakdown of work and cost structures for an long-term digital preservation system that incorporates cloud computing technologies and satisfies all the requirements set by businesses.

Combining breakdowns and cost drivers resulted in cost equations that would always estimate a single point cost for the long-term digital preservation systems following the same requirements and the main OAIS standard. It was noted that the OAIS standard was not followed by all previous cost models; they follow its recommendations more or less faithfully. This research project worked close to OAIS not only to keep its updates and upgrades easier for future users, but also to refer less experienced users to a very detailed standard.

7.5 Conclusion

This research thesis succeeded in meeting all the seven objectives set in Chapter 1. An understanding of all the domains of research, from long-term digital preservation to cost modelling, was reached. The research has provided evidence from which to understand uncertainty cost estimation and definitions and the relationship of cost estimation to obsolescence. It provides a simple yet

clear list of the different sector requirements to make it easy for users afterwards to follow. A standard obligatory lifecycle has been developed that allows authentic cost drivers and work and cost breakdown structures to be constructed. Additionally, the research has been validated by industry experts, alongside a real long-term digital preservation system design exercise for real-life use.

Studying all the potential cost estimation techniques, Activity Based Costing was selected to stabilise a reliable, simple and upgradable initial cost model for this specific purpose, while maintaining the ability to change to a more sophisticated estimation technique. The production of a single-point cost estimate at this stage meant that the research output could start to produce some results at an early stage of the research, which in turn gave more time to validate all the previous stages to the output with a great many experts.

An uncertainty identification process was designed; it measured the impact of uncertainties on cost drivers, while also developing a taxonomy of obsolescence and measuring its impact on cost drivers (6). Finally, the impacts of both uncertainties and obsolescence issues with a single-point estimate were integrated to generate a three-point estimate. All these outputs were using in constructing a viable, validated and concept-proof cost modelling framework (7).

The whole framework, with all its processes, inputs and outputs were thoroughly validated with several industry experts. With compliments on the design details, accuracy and ingenuity, the framework was deemed by experts to be matching current best practices and representing real-life LTDP systems, while taking into consideration its compatibility with the potential of other business sectors.

7.6 Research Limitations and Future Work

The limitations of this research project are mainly generated from the subjectivity of some areas that depended on experts' knowledge and experience. The author refrained from influencing any data collection sessions, by adopting different setups in different sessions. One-to-one interviews telephone interviews and group interviews were employed to reduce any influence of this kind and to encourage experts both by accepting solitary ideas and brain storming.

Case studies were developed with the understanding of current experts and their prediction of future costs; while these were reasonably close to the calculated costs, collating future data and comparing it with cost estimates is a good means of making sure that any deviation in the framework can be found that could develop from the advent of a new unknown.

There is a need for further research work to build on the results found by this project. Future work is what ensures that more benefit can be reaped from the research output, and what keeps the area of study up to date with scientific developments in all the aspects of the research. The following are some of the possible research studies that could be carried out:

- Integrating cost models with big data analytics on all the preserved information for a business sector, such as healthcare.
- A thorough compatibility study to assess in more depth whether or not the framework could be adapted to all business sectors. Long-term historical cost data would help this study, especially with validation.
- Different cost estimation techniques could be investigated, collating more cost values over a long period.

- Uncertainties detected by the framework might have some cross-impact on each other, or on some obsolescence issues or by them. Detailed study of the cross-impact of uncertainties can add to knowledge.
- A study of implementing internet of things in long-term digital preservation systems could be carried out, where data is submitted from devices and sensors, etc. directly to a preservation system. A look at the cost, uncertainties and implications of this should prove very useful for moving forward.
- Studying the possibility of streamlining the framework and bringing it more into line with lean principles.

REFERENCES

AACEI (2007), *AACE International Recommended Practice No. 10S-90: COST ENGINEERING TERMINOLOGY*, AACE International, Inc., West Virginia, USA.

Amazon AWS (2014), *How AWS Pricing Works*, available at: http://media.amazonwebservices.com/AWS_Pricing_Overview.pdf (accessed 13/08/014).

Anbu, J. and Chibambo, M. (2009), "Digital Preservation: Issues and Challenges.", *Trends in Information Management,* vol. 5, no. 1.

Andrade, M. C., Pessanha Filho, R. C., Espozel, A. M., Maia, L. O. A. and Qassim, R. Y. (1999), "Activity-based costing for production learning", *International Journal of Production Economics,* vol. 62, no. 3, pp. 175-180.

Ayris, P., Davies, R., McLeod, R., Miao, R., Shenton, H. and Wheatley, P. (2008), "The LIFE2 final project report"

Baker, M., Shah, M., Rosenthal, D. S., Roussopoulos, M., Maniatis, P., Giuli, T. J. and Bungale, P. (2006), "A freshlook at the reliability of long-term digital storage", ACM SIGOPS Operating Systems Review, Vol. 40, ACM, pp.221

Ball, A. (2008), "Report from the PV 2007 Conference, DLR Oberpfaffenhofen, October 9–11, 2007", *International Journal of Digital Curation,* vol. 2, no. 2, pp. 69-81.

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

Beagrie, N., Lavoie, B. and Woollard, M. (2010), *Keeping Research Data Safe 2,* , Charles Beagrie Limited, UK.

Bhushan, N. and Rai, K. (2004), *Strategic decision making: applying the analytic hierarchy process,* Springer Verlag, United States of America.

Bide, M., Potter, J. and Watkinson, A. (1999), "Digital preservation: an introduction to the standards issues surrounding the deposit of non-print publications", *Book Industry Communication*

Bode, J. (2000), "Neural networks for cost estimation: simulations and pilot application", *International Journal of Production Research,* vol. 38, no. 6, pp. 1231-1254.

Boehm, B., Abts, C. and Chulani, S. (2000), "Software development cost estimation approaches—A survey", *Annals of Software Engineering,* vol. 10, no. 1, pp. 177-205.

Boehm, B. W. (1984), "Software engineering economics", *Software Engineering, IEEE Transactions on*, no. 1, pp. 4-21.

Bonabeau, E. (2002), "Agent-based modeling: Methods and techniques for simulating human systems", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 99, no. Suppl 3, pp. 7280.

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

Brians, C. L., Willnat, L., Manheim, J. and Rich, R. (2016), Empirical political analysis, 8th ed, Routledge, Abingdon, Oxon

Brimson, J. A. (1998), "Feature costing: beyond ABC", *Journal of cost management,* vol. 12, no. http://maaw.info/ArticleSummaries/ArtSumBrimson1998.htm (Last visited on 29/07/2011), pp. 6-13.

British Library (2012), *17th and 18th Century Burney Collection Database*, available at: http://www.bl.uk/reshelp/findhelprestype/news/newspdigproj/burney/index.html (ac cessed February 5th 2012).

British Library (2007), *British Library Digital Preservation Strategy*, British Library.

Brown, A. (2008), "Digital Preservation Guidance Note 1: Selecting file formats for long-term preservation", *London: The National Archives,* vol. 5.

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).

Chapman, S. (2006), "Counting the costs of digital preservation: is repository storage affordable?", *Journal of Digital Information,* vol. 4, no. 2.

Charles Beagrie Limited (2010), *Ensuring Perpetual Access: Establishing a Federated Strategy on Perpetual Access and Hosting of Electronic Resources for Germany*, Charles Beagrie Limited in association with Globale Informationstechnik GmbH, Bonn.

Chuku, G. (2012), *A FRAMEWORK FOR HANDLING UNCERTAINTIES IN LONG TERM DIGITAL PRESERVATION* (MSc thesis), Cranfield University, Cranfield.

CMDP (2005), *Costs of Digital Preservation,* Nationaal Archief, The Hague, Netherlands.

CMDP 2 (2011), *Costs of Digital Preservation - Project Report for Phase 2,* The Danish Royal Library and The Danish National Archives, Denmark

Creswell, J. (2003). Research design: Qualitative, Quantitative, and Mixed Methods Approaches. London: Sage

Creswell, J. W. and Poth, C. N. (2017), Qualitative inquiry and research design: Choosing among five approaches, 4th ed, Sage publications, London

Day, M. (2006), "The long-term preservation of Web content", Web archiving, pp. 177-199

DCC and Digital Curation Centre (2010c), *Trustworthy Repositories*, available at: http://www.dcc.ac.uk/resources/tools-and-applications/trustworthy-repositories (accessed 12/07/2011).

DCC and Digital Curation Centre (2010b), *Trust through self assessment*, available at: http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/trust-through-self-audit (accessed 12/07/2011).

DCC and Digital Curation Centre (2010a), *Nestor Catalogue of Criteria for Trusted Digital Repositories*, available at: http://www.dcc.ac.uk/resources/tools-and-applications/nestor (accessed 02/08/2011).

DPC (2012), *Introduction - Definitions and Concepts*, available at: http://www.dpconline.org/advice/preservationhandbook/introduction/definitions-and-concepts (accessed 12/01).

Duverlie, P. and Castelain, J. M. (1999), "Cost estimation during design step: parametric method versus case based reasoning method", *The international journal of advanced manufacturing technology,* vol. 15, no. 12, pp. 895-906.

Edalew, K. O., Abdalla, H. S. and Nash, R. J. (2001), "A computer-based intelligent system for automatic tool selection", *Materials & Design,* vol. 22, no. 5, pp. 337-351.

ENSURE (2012), *Activity II Scientific Report,* D20.1.A, ENSURE, Brussels.

ENSURE (2010), *Grant agreement for the Collaborative project: "Enabling kNowledge Sustainability Usability and Recovery for Economic value"*, Annex1 - Description of work. ed., Seventh Framework Programme, European Union.

Epstein, J. M. and Axtell, R. (1996), *Growing artificial societies: social science from the bottom up,* The MIT Press.

Erkoyuncu, J., Roy, R., Shehab, E. and Wardle, P. (2009), "Uncertainty challenges in service cost estimation for product-service systems in the aerospace and defence industries", *Proceedings of the 19th CIRP Design Conference–Competitive Design,* Cranfield University Press.

Erkoyuncu, J. A. (2011), *Cost uncertainty management and modelling for industrial product-service systems* (PhD thesis), Cranfield University, Cranfield.

EU Commission (2006), "Directive 2006/24/EC of the European parliament and of the council of 15 march 2006'", Office Journal of the European Union, vol. 105, pp. 54, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:105:0054:0063:EN:PDF (Last accessed 25/08/2015) .

Evans, D. K., Lanham, J. D. and Marsh, R. (2005), "Cost Estimation Method Selection: Matching User Requirements and Knowledge Availability to Methods", *Systems Engineering and Estimation for Decision Support (SEEDS) Group, University of West of England, Bristol, UK*.

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: Modeling and Implementation in Preservation Aware Storage.", *Workshop on the Theory and Practice of Provenance*

Foster, I., Zhao, Y., Raicu, I. and Lu, S. (2008), "Cloud computing and grid computing 360-degree compared", *Grid Computing Environments Workshop, 2008. GCE'08,* Ieee, pp. 1.

Gayretli, A. and Abdalla, H. (1999), "An object-oriented constraints-based system for concurrent product development", *Robotics and computer Integrated Manufacturing,* vol. 15, pp. 133-144.

Gayretli, A. and Abdalla, H. S. (1999), "A feature-based prototype system for the evaluation and optimisation of manufacturing processes", *Computers & Industrial Engineering,* vol. 37, no. 1-2, pp. 481-484.

Giaretta, D. (2011), Advanced Digital Preservation, Springer, Berlin

Gong, C., Liu, J., Zhang, Q., Chen, H. and Gong, Z. (2010), "The characteristics of cloud computing", *Parallel Processing Workshops (ICPPW), 2010 39th International Conference on,* 13 - 16 September, San Diego, CA, USA, IEEE, pp. 275 - 279.

Google (2011)*, Power Distribution System for Google Data Centre Canada*, available at: http://www.google.ca/corporate/datacenter/images/schematic.gif (accessed 07/11/2014).

Graham, P. S. (1993), "Intellectual preservation in the electronic environment", submitted to be part of Arnold Hirshon, editor. "After the Electronic Revolution, Will You Be the First to Go"

Granger, S. (2000), "Emulation as a digital preservation strategy", D-Lib Magazine, vol. 6, no. 10 (http://www.dlib.org/dlib/october00/granger/10granger.html) (Last accessed 30/08/2017)

Greenberg, A., Hamilton, J., Maltz, D. A. and Patel, P. (2008), "The cost of a cloud: research problems in data center networks", *ACM SIGCOMM Computer Communication Review,* vol. 39, no. 1, pp. 68-73.

Hedstrom, M. (1997), "Digital preservation: a time bomb for digital libraries", *Computers and the Humanities,* vol. 31, no. 3, pp. 189-202.

Hedstrom, M. (1997), "Digital preservation: a time bomb for digital libraries", *Computers and the Humanities,* vol. 31, no. 3, pp. 189-202.

Helmer, O. (1967), "Analysis of the future: The Delphi method", *Analysis of the future*.

Hendley, T. (1998), "Comparison of methods and costs of digital preservation", *British Library Research and Innovation Report 106,* British Library Research and Innovation Centre, West Yorkshire*,* Citeseer, 121 pp

Higgins, S. (2009), *PREMIS Data Dictionary for Preservation Metadata,* http://hdl.handle.net/1842/3339, Digital Curation Centre.

Hofman, H., (2009), *Digital Preservation Process: Preparation and Requirements*, Nationaal Archief Netherlands (http://www.planetsproject. eu/docs/presentations/Hofman_DPProcess.pdf) (Last visited 25/06/2011), Barcelona.

Holdsworth, D. (2001), "C-ing ahead for digital longevity", available at: http://sw.ccs.bcs.org/CAMiLEON/dh/cingahd.html (accessed 07/01/2013)

Holdsworth, D. (2006), Digital Preservation, pp 32-59, Facet Publishing, London

Hole, B., Wheatley, P., Lin, L., McCann, P. and Aitken, B. (2010), "The Life3 Predictive Costing Tool for Digital Collections", *New Review of Information Networking,* vol. 15, no. 2, pp. 81-93.

Hubbard, D. W. (2010), *How to measure anything: Finding the value of intangibles in business,* Wiley.

Hughes, L. M. and GREN, D. (2004), *Digitizing collections: strategic issues for the information manager,* Facet Publishing, London.

Hughes, L. (2002), *Digitizing collections: strategic issues for the information manager,* Facet, London.

Hundal, M. (1993), "Design to cost", *Concurrent Engineering: Contemporary Issues and Modern Design Tools, Chapman and Hall,* pp. 330-351.

Hunolt, G., Booth, B. and Banks, M. (2008), *Cost Estimation Toolkit Technical Description Document,* 2.4, NASA.

IP (2010)*, Digital Preservation*, available at: http://ip.org.au/digital-preservation/ (accessed 11/23/2012).

Jung, J. Y. (2002), "Manufacturing cost estimation for machined parts based on manufacturing features", *Journal of Intelligent Manufacturing,* vol. 13, no. 4, pp. 227-238.

Kejser, U. B. (2009), *Modelling the Cost and Quality of Preservation Imaging and Archiving* (PhD thesis), Danish School of Conservation and The Danish Research School of Cultural Heritage, Denmark.

Kejser, U. B., Nielsen, A. B. and Thirifays, A. (2009), "Cost Model for Digital Curation: Cost of Digital Migration", *iPRES 2009: The Sixth International Conference on Preservation of Digital Objects,* Vol. 6th, San Francisco, California, California, pp. 98.

Kingsman, B. G. and De Souza, A. A. (1997), "A knowledge-based decision support system for cost estimation and pricing decisions in versatile manufacturing companies", *International Journal of Production Economics,* vol. 53, no. 2, pp. 119-139.

Kirchhoff, A. J. (2008), "Digital preservation: challenges and implementation", *Learned Publishing,* vol. 21, no. 4, pp. 285-294.

Kulovits, H. and Raube, A. (2009), *Digital Preservation in Radiology. Ensuring long-term accessibility of digital medical images,* Digital preservation in Europe, Vienna, Austria

Kumar, R. (2005), "*Research Methodology: A Step-by-Step Guide for Beginners*", Sage, 2005

Kuny, T. (1998), "*The digital dark ages? Challenges in the preservation of electronic information*", *International Preservation News*, pp. 8-13.

Lawrence, G. W., Kehoe, W. R., Rieger, O. Y., Walters, W. H. and Kenney, A. R. (2000), Risk Management of Digital Information: A File Format Investigation. ERIC

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.

LIFE (2010)*, LIFE^3: An Overview*, available at: http://www.life.ac.uk/3/ (accessed September 15th 2012).

LIFE (2008)*, LIFE^2: An Overview*, available at: http://www.life.ac.uk/2/ (accessed September 29th 2011).

LIFE (2007)*, What is LIFE: Life Cycle Information for E-Literature*, available at: http://www.life.ac.uk/about/ (accessed September 15th 2011).

Likert, R. (1932), "A technique for the measurement of attitudes.", *Archives of psychology*, no. 140.

Ling, D. (2005), *Railway Renewal and Maintenance Cost Estimating*Cranfield University, Cranfield, United Kingdom.

Marcum, D. (1997), "A moral and legal obligation: Preservation in the digital age"

McLeod, R., Wheatley, P. and Ayris, P. (2006), "Lifecycle information for e-literature: full report from the LIFE project"

McManus, H. and Hastings, D. (2004), "A Framework for Understanding Uncertainty and its Mitigation and Exploitation in Complex Systems", *INCOSE International Symposium,* Vol. 15, July, NY, Wiley Online Library, Rochester, pp. 1-20.

Mell, P. and Grance, T. (2009), "The NIST definition of cloud computing", *National Institute of Standards and Technology,* vol. 53, no. 6, pp. 50.

Microsoft (2014)*, Support for Windows XP ended April 8th, 2014*, available at: https://www.microsoft.com/en-us/WindowsForBusiness/end-of-xp-support (accessed 06/06/2014).

Microsoft (2008)*, Datacenter Architecture for Environmental Sustainability – "Green Datacenters"*, available at: http://blogs.technet.com/b/nymciblog/archive/2008/03/21/datacenter-architecture-for-environmental-sustainability-green-datacenters.aspx (accessed 25/10/2014)

Miles, M. B., Huberman, A. M. and Saldana, J. (2014), Qualitative data analysis: A method sourcebook, 3rd ed, Sage Publications, California, USA, P 20

Ministry of Defence (MoD) (2007)*, Three Point Estimates and Quantitative Risk Analysis: A Process Guide for Risk Practitioners (Unclassified)*, available at: https://www.aof.mod.uk/aofcontent/tactical/risk/downloads/3pepracgude.pdf (accessed 18/03/2014).

NASA (2011)*, Cost Estimation Toolkit (CET) Software Package*, available at: http://opensource.gsfc.nasa.gov/projects/CET/ (accessed January 18th 2012).

NASA (1996), *Parametric Cost Estimating Handbook,* 1st ed, Nasa, http://cost.jsc.nasa.gov/pcehg.html (Last visited 01/08/2011).

NATO (2007)*, Methods and Models for Life Cycle Costing*, available at: https://www.cso.nato.int/pubs/rdp.asp?RDP=RTO-TR-SAS-054 (accessed 14/12/2014).

Neervens, A. (2009), "The Battle Against Digital Obsolescence: Exploring Strategies of Digital Preservation in New Media and New Media Art", *New Media Theories*, pp. 1

Newton, S. (2009), "A Critique of Initial Budget Estimating Practice", Anita Cerić and Mladen Radujković (ed.), in: *Proceedings of Construction Facing Worldwide Challenges Conference: CIB Joint International Symposium W55/W65,* 27/09/2009 - 01/10/2009, Dubrovnik, Croatia, International Council for Research and Innovation in Building and Construction, Croatia, pp. 271-280.

Niazi, A., Dai, J. S., Balabani, S. and Seneviratne, L. (2006), "Product cost estimation: Technique classification and methodology review", *Journal of manufacturing science and engineering,* vol. 128, pp. 563.

NLOA and National Library of Australia (1999), *Preserving Access to Digital Information - Encapsulation*, available at: http://www.nla.gov.au/padi/topics/20.html (accessed 02/07/2011).

Pallis, G. (2010), "Cloud Computing", *IEEE INTERNET COMPUTING,* vol. 10, pp. 70-73.

Petrovic, D. (2011), "Why Digital Preservation is Important for Everyone", Australian Science

Purdy, G. (2010), "ISO 31000: 2009—Setting a New Standard for Risk Management", *Risk Analysis,* vol. 30, no. 6, pp. 881-886.

Rehman, S. and Guenov, M. D. (1998), "A methodology for modelling manufacturing costs at conceptual design", *Computers & Industrial Engineering,* vol. 35, no. 3-4, pp. 623-626.

Riley, J., (2007), "*An Introduction to PREMIS",* (unpublished Presentation), https://scholarworks.iu.edu/dspace/bitstream/handle/2022/16507/premis.pdf?sequence=2&isAllowed=y (Last visited 08/08/2011), Indiana

Robson, C. (2002), Real world research: A resource for social scientists and practitioner-researchers, 2nd ed, Wiley-Blackwell, United Kingdom

Romero Rojo, F. J. (2011), "Development of a framework for obsolescence resolution cost estimation", (PhD thesis), Cranfield University, Cranfield, Bedfordshire

Rosenthal, A., Mork, P., Li, M. H., Stanford, J., Koester, D. and Reynolds, P. (2010), "Cloud computing: a new business paradigm for biomedical information sharing", Journal of Biomedical Informatics, vol. 43, no. 2, pp.342-353

Rothenberg, J. (1995), "Ensuring the longevity of digital documents", Scientific American, vol. 272, no. 1, pp. 42-47

Rothenberg, J. (1999), *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation. A Report to the Council on Library and Information Resources.* Council on Library and Information Resources, 1755 Massachusetts Ave., NW, Washington, DC 20036.

Rothenberg, J. (2000), *An Experiment in Using Emulation to Preserve Digital Publications,* Koninklijke Bibliotheek and RAND-Europe, Netherlands.

Roy, R. (2003), *Decision Engineering Report Series "Cost engineering: why, what and how?",* ISBN 1-861940-96-3, Cranfield University, Cranfield.

Roy, R., Kelvesjo, S., Forsberg, S. and Rush, C. (2001), "Quantitative and qualitative cost estimating for engineering design", *Journal of Engineering Design,* vol. 12, no. 2, pp. 147-162.

Roy, R. and Erkoyuncu, J. A. (2011), "Service cost estimation challenges in industrial product-service systems", in *Functional Thinking for Value Creation,* Springer, pp. 1-10.

Rush, C. and Roy, R. (2001), "Expert judgement in cost estimating: Modelling the reasoning process", *Concurrent Engineering,* vol. 9, no. 4, pp. 271.

Russell, K. (2000), "Digital preservation and the CEDARS project experience", *New Review of Academic Librarianship,* vol. 6, pp. 139-154.

Ruusalepp, R. (2003), *AHDS Digital Preservation Glossary,* , Estonian Business Archives, Ltd.

Sandborn, P. (2007), "Software obsolescence–complicating the part and technology obsolescence management problem", *IEEE Transactions on Components and Packaging Technologies,* vol. 30, no. 4, pp. 886–888.

Sandborn, P. and Singh, P. (2004), "Forecasting Technology Insertion Concurrent with Design Refresh Planning for COTS-Based Electronic Systems", *Spectrum,* vol. 38, no. 2, pp. 25-28.

Sandborn, P., Jung, R., Wong, R. and Becker, J. (2007), "A taxonomy and evaluation criteria for DMSMS tools, databases and services", *Proceedings of the Aging Aircraft Conference,* Palm Springs, California, .

Sandborn, P. (2013), "Design for obsolescence risk management", *Procedia CIRP,* vol. 11, pp. 15-22.

Sandborn, P., Prabhakar, V. J. and Kusimo, A. (2012), "Modeling the obsolescence of critical human skills necessary for supporting legacy systems", *ASME 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference,* American Society of Mechanical Engineers, pp. 723.

Shehab, E. and Abdalla, H. (2002), "An intelligent knowledge-based system for product cost modelling", *The international journal of advanced manufacturing technology,* vol. 19, no. 1, pp. 49-65.

Shehab, E. M. and Abdalla, H. S. (2002), "A design to cost system for innovative product development", *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture,* vol. 216, no. 7, pp. 999-1019.

Shehab, E. and Abdalla, H. (2001), "Manufacturing cost modelling for concurrent product development", *Robotics and Computer-Integrated Manufacturing,* vol. 17, no. 4, pp. 341-353.

Shehab, E., Chuku, G. and Badawy, M. (2013), "A FRAMEWORK FOR IDENTIFYING UNCERTAINTIES IN LONG-TERM DIGITAL PRESERVATION", *Proceedings of the 11th International Conference on Manufacturing Research ICMR 2013,* 19-20/09/2013, Cranfield, Cranfield University, Cranfield

Shepard, T (1998), "Universal Preservation Format (UPF): conceptual framework" RLG DigiNews, 2(6)

Shepperd, M. and Cartwright, M. (2001), "Predicting with sparse data", *Software Engineering, IEEE Transactions on,* vol. 27, no. 11, pp. 987-998.

Shermon, D. (ed.) (2009), *Systems Cost Engineering: Program Affordability Management and Cost Control,* Surrey, England.

Shields, P. and Rangarajan, N. (2013), A Playbook for Research Methods: Integrating Conceptual Frameworks and Project Management, New Forums Press, Stillwater, Oklahoma

Singh, P. and Sandborn, P. (2006), "Obsolescence driven design refresh planning for sustainment-dominated systems", *The Engineering Economist,* vol. 51, no. 2, pp. 115-139.

Smith, E. P. (2002), "Uncertainty Analysis", *Encyclopedia of environmetrics,* vol. 4, pp. 2283-2297.

Solomon, R., Sandborn, P. A. and Pecht, M. G. (2000), "Electronic part life cycle concepts and obsolescence forecasting", *Components and Packaging Technologies, IEEE Transactions on,* vol. 23, no. 4, pp. 707-717.

Stanger, N. (2011), "Keeping research data safe", *Computer and Information Science Seminar Series*

Tannert, C., Elvers, H. D. and Jandrig, B. (2007), "The ethics of uncertainty. In the light of possible dangers, research becomes a moral duty", *EMBO reports,* vol. 8, no. 10, pp. 892.

Thunnissen, D. P. (2005), *Propagating and Mitigating Uncertainty in the Design of Complex Multidisciplinary Systems* (unpublished PhD thesis), California Institute of Technology, Pasadena, California.

Time, R. T. "Researching long term digital preservation approaches in the dutch digital preservation testbed (testbed digitale bewaring)", .

TRC and Trusted Repositiories Certification (2006), "Catalogue of Criteria for Trusted Digital Repositories", vol. 1, pp. 1-39.

United Kingdom Parliament, (2014), The Data Retention Regulations, https://www.legislation.gov.uk/uksi/2014/2042/pdfs/uksi_20142042_en.pdf (last accessed 25/08/2014), No. 2042, London

Venkatachalam, A., Mellichamp, J. M. and Miller, D. M. (1993), "A knowledge-based approach to design for manufacturability", *Journal of Intelligent Manufacturing,* vol. 4, no. 5, pp. 355-366.

Walker, W. E., Harremoës, P., Rotmans, J., van der Sluijs, Jeroen P, van Asselt, M. B., Janssen, P. and Krayer von Krauss, Martin P (2003), "Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support", *Integrated assessment,* vol. 4, no. 1, pp. 5-17.

Waters, D. and Garrett, J. (1996), *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*, The Commission on Preservation and Access and The Research Libraries Group, Washington DC and Mountain View CA, 64 pp.

Waugh, A., Wilkinson, R., Hills, B. and Dell'oro, J. (2000), "Preserving Digital Information Forever"

Wheatley, P., Ayris, P., Davies, R., McLeod, R. and Shenton, H. (2007), "LIFE: costing the digital preservation lifecycle", *iPRES 2007,* October 2007, China

Wheatley, P. and Hole, B. (2009), "LIFE3: Predicting Long Term Digital Preservation Costs", *iPRES 2009: The Sixth International Conference on the Preservationof Digital Objects,* Vol. 6th, October 5 - 6 2009, San Francisco, California, California Digital Library, California, pp. 206.

Wierda, L. S. (1991), "Linking design, process planning and cost information by feature-based modelling.", *J.ENG.DES.,* vol. 2, no. 1, pp. 3-19.

Wu, L. (1997)*, The comparison of the software cost estimating methods*, available at: http://www.compapp.dcu.ie/~renaat/ca421/LWu1.html (accessed April 5th 2012).

Xu, Y., Elgh, F., Erkoyuncu, J. A., Bankole, O., Goh, Y., Cheung, W. M., Baguley, P., Wang, Q., Arundachawat, P. and Shehab, E. (2012), "Cost Engineering for manufacturing:

Current and future research", *International Journal of Computer Integrated Manufacturing,* vol. 25, no. 4-5, pp. 300-314.

Xue, P., Badawy, M., Shehab, E. and Baguley, P. (2011), "Cost Modelling for Long-Term Digital Preservation: Challenges and Issues", *9th International Conference on Manufacturing Research ICMR 2011,* 6th of September 2011, Glasgow.

Youseff, L., Butrico, M. and Da Silva, D. (2008), "Toward a unified ontology of cloud computing", *Grid Computing Environments Workshop, 2008. GCE'08,* IEEE, pp. 1-10.

Zhang, Q., Cheng, L. and Boutaba, R. (2010), "Cloud computing: state-of-the-art and research challenges", *Journal of internet services and applications,* vol. 1, no. 1, pp. 7-18.

Zissis, D. and Lekkas, D. (2012), "Addressing cloud computing security issues", *Future Generation Computer Systems,* vol. 28, no. 3, pp. 583-592.

APPENDICES

The following appendices adds some related material to the work done in this research project, which should enhance the readers understanding of what work was carried.

Appendix A Questionnaires

A.1 Long-Term Digital Preservation Current State of the Art Questionnaire

The purpose of this questionnaire is to capture the AS-IS of the long-term digital preservation activities performed by the Company X.

General Questions:

1. Name:
2. Job Title:
3. Job Role:
4. Company/Department Structure:
5. Experience:
    a. In this company: months/years
    b. In Digital Preservation: months/years % of time.
6. Do you have other activities in the company:
7. Which levels of preservation is employed by your company?

☐ We don't preserve ☐ Just a copy of the files

☐ A data archive and Storage

☐ Fixity check ☐ Migration ☐ Emulation

☐ Data description added ☐Other

8. Do you have a special preservation team? ☐ Yes ☐No
    a. If yes, how many employees are there in this team? / % of time.

b. If no, who is doing the preservation for you? And how many are involved?

9. What is the average overhead cost of employees in your company, including salaries? (Overhead costs e.g. pension, taxes, insurance, etc.) [if overhead costs are unknown please mention the average salary for an employee]
10. What is the whole lifecycle process of digital preservation adopted by the company?
11. How did your company used to carry out data preservation before these processes were activated?
12. How did you move to accommodate the present preservation processes over the previous processes?
13. What is your future preservation plan?

Data Generation:

14. How many companies/labs/departments are generating data to be preserved:

15. Who defines the preservation policies for the data? Position
16. What is the volume of the data generated daily?
17. What is the future anticipated change in the volume of data generated?
18. What data types are generated (e.g. documents, images, etc…)?
19. What are the file formats of each data type? File format generated and the file format preserved; if they are different (please fill in following table).
20. What size contribution does each file format has from the total data volume? (please fill in following table)

Data Type	Generation format	Preservation format	% Size of total

21. What are the activities carried-out to ingest this data?
22. Who generates the meta-data for the data to be preserved?
23. Can you quantify, roughly, the percentage of the size of the meta-data required for the data to be preserved?
24. For how long are your preservation plans designed for?

Preservation

25. What are the activities involved in the preservation phase?
26. Do you follow any specific standard while carrying out these activities? (e.g. OAIS, TRAC)
27. Do you need to preserve any special software or environment for these files?
28. What are your current infrastructure resources available for the preservation activities? Please mention any known:
    a. Cost of the required Physical space (e.g. building hire, construction costs,)
    b. Costs of Hardware (e.g. servers, racks, storage)
    c. Costs of Software (e.g. operating systems, database licencing,)
    d. Costs of Security (e.g. physical or software)

e. Costs of Utilities requirements (e.g. power, cooling)

f. Costs of Staff (e.g. manager, technician, technology watch)

g. Other special infrastructure costs

29. How often do you perform fixity checks on your preserved data?

30. How long does it take to do this for a set unit of data?

31. Are the fixity checks manual or automated? Any known costs?

32. What is the time required for the fixity checks?

Access

33. What are the activities involved in accessing the preserved data?

34. Who is entitled to access the preserved files? (Please fill in the following table)

35. What is the expected access rate? (Please fill in the following table)

Access Allowed to	File Type	Expected Access Rate

36. How are the files delivered to the user?

37. Are they allowed to edit the accessed files?

38. What are the required resources to perform access activities?

a. Physical space

b. Hardware

c. Software

d. Staff

e. Other resources

39. Does your company have any previous cloud storage experience?

Uncertainties:

Uncertainties in cost estimation for LTDP can be defined as "*unexpected occurrence and/or the lack of enough knowledge about the processes carried-out, where this will generate noise in the cost estimated*".

40. What are the main uncertainties that face each preservation activity?

Uncertainty	Preservation Phase	Frequency

41. What prioritisation category do you think is suitable to differentiate between these uncertainties?

42. What are the main strategies that you adopt to overcome these uncertainties?

43. What are the main obsolescence issues that face your preservation process?

Obsolescence	Preservation Phase	Mitigation

44. What prioritisation category do you think is suitable to differentiate between these obsolescence issues?
45. What are the main strategies that you have adapted to overcome these obsolescence issues?

A.2 Sector Differences Questionnaire

The purpose of this questionnaire is to capture the AS-IS and the Sector Differences of for the long-term digital preservation requirements for NHS.

General Questions:

1. Name:
2. Job Title:
3. Job Role:
4. Department Structure:

5. Experience:
    c. In this company: months/years
    d. In IT in general: months/years
    e. In Digital Preservation (if any): months/years % of time.

Digital Preservation:

6. Which levels of preservation is employed by your company?

☐ Just a copy of the files ☐ A data archive and Storage ☐ Fixity check

☐ Migration ☐ Emulation ☐ Encapsulation

☐ Museum Approach ☐ Data description added ☐ other……

7. Do you have a special preservation team? ☐ Yes ☐ No (*if No, please go to question 9*)

   a. If yes, how many employees are in this team? / % of their time

   b. If no, who is carrying-out those preservation activities for you?

   c. And how many are involved?

8. What is the whole lifecycle process of digital preservation adopted by the company?

File Types

9. What are the basic file types generated by your company? (*may chose more than one*)

☐ Images ☐ Video ☐ Alphanumeric ☐ other

10. For these file types what are the file formats? And what is the volume of each format of your total data generation?

File Type					
File Format					
GB/Month					

11. Do you know any costs incurred to generate each file format per gigabyte (GB)? *If no please go to question 12*

File Format					
Cost/GB					

12. Do you consider any file formats to be complex? ☐ Yes ☐ No (*if no go to question 14*)

13. Can you rate the file complexity with comparison to other formats? 1 is least complex and 5 is the highest complexity (*if no complexity rating go to question 14*)

File Format					
Complexity Rating					

Why?					

14. What are the problems that could be face with each file format? (*if non go to question 16*)

File Format				
Problem				

15. How do you ensure these problems are avoided?

Problem				
Avoidance Strategy				

16. How long are you planning to preserve each file format?

File Format					
Period					

Policy:

17. Do you have an existing data preservation policy? ☐Yes ☐No (*if no go to question 20*)

18. Could this plan be supplied to Cranfield University's ENSURE research team? ☐Yes ☐No

19. Could you summarise your data preservation policy?

20. What are the legal obligations and requirements for digital preservation acted on your company?

21. Are there any copyright issues that could affect your information targeted for digital preservation? ☐ Yes ☐ No (If yes what are these issues - *if no go to question 23*)

22. Do these copyright issues generate any cost that is incurred by your company?

☐ Yes ☐ No (If yes please state – *if No go to question 23*)

23. Who generates the metadata for the data to be preserved?

24. How much metadata, in GB, is required for one GB of each file format? (*if unknown go to question 25*)

File Format					
Metadata size (GB)					

Access

25. What is the expected access rate annually?

File Format					
Annual Access Rate					

26. Do you expect this rate to change overtime? If yes by what percentage?

27. What are the main risks that face your preserved information from your point of view?

A.3 Delphi Workshop

Name: ……

Company: ……

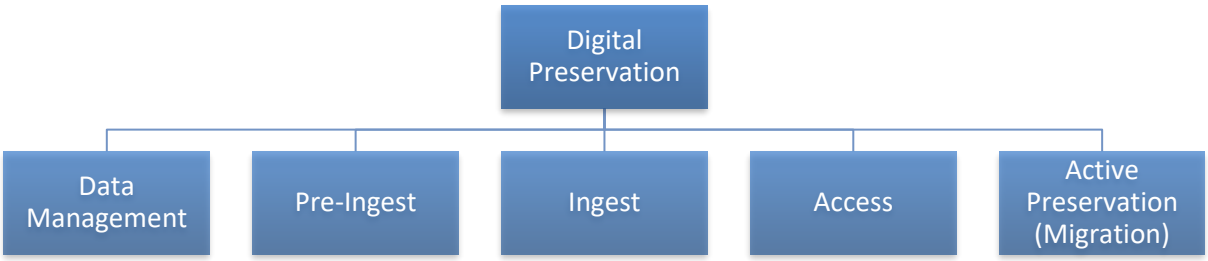
Job Title: ……

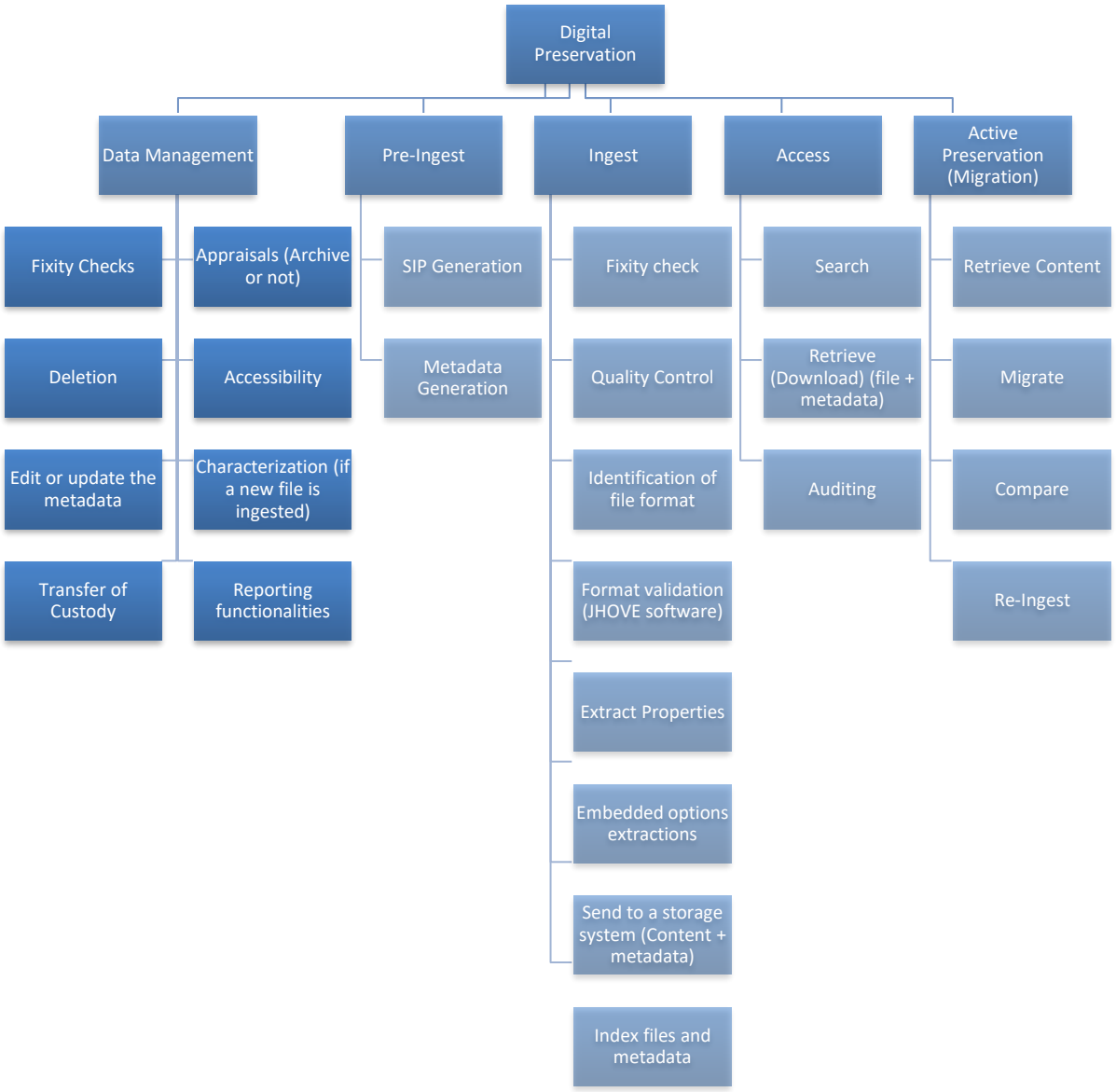
Sector of Expertise:

☐ Clinical Trails ☐ Financial ☐ Healthcare

☐ Science Facility

The following Figures 1 and 2 show the whole lifecycle Work Breakdown structure of Long-Term Digital Preservation





1. What are the files generated and the access rate expected/year?

File type	File format	Size generated/year (GB)	Expected access rate/year/file type (numbers)

LTDP:

2. What are the known costs endured with each activity

Activity	Cost
Data Management	
Pre-Ingest	
Ingest	
Access	
Active Preservation	

3. What are your current infrastructure resources available for the total lifecycle of preservation activities? Please mention any known:

Category	Cost
Physical space (e.g. building hire, construction costs)	
Costs of Hardware (e.g. servers, racks, storage)	
Costs of Software (e.g. operating systems, database licencing)	
Costs of Security (e.g. physical or software)	
Utilities requirements (e.g. power, cooling)	
Costs of Staff (e.g. manager, technician, technology watch)	
Other costs	

4. How often do you perform fixity checks on your preserved data? ……

5. What unit do you prefer to use when measuring fixity check? ☐ Giga Bytes ☐ No. Files

6. How long does it take to do this set unit? ……

7. Does your company have any previous cloud storage experience? ☐ Yes ☐ No

8. If yes, was it only storage or did it include computing?

   ☐ Storage ☐ Computing ☐ Both

9. Can you give information on usage, data size, costs and any suitable information of each of the cloud services your company had?

    a. Storage ……

    b. Computing ……

    c. Both ……

Uncertainties:

Uncertainties in cost estimation for LTDP can be defined as "*unexpected occurrence and/or the lack of enough knowledge about the processes carried-out, where this will generate noise in the cost estimated*".

10. What are the main uncertainties that face each preservation activity?

Uncertainty	Mitigation Strategy	Cost

11. What prioritisation category do you think is suitable to differentiate between these uncertainties? ……

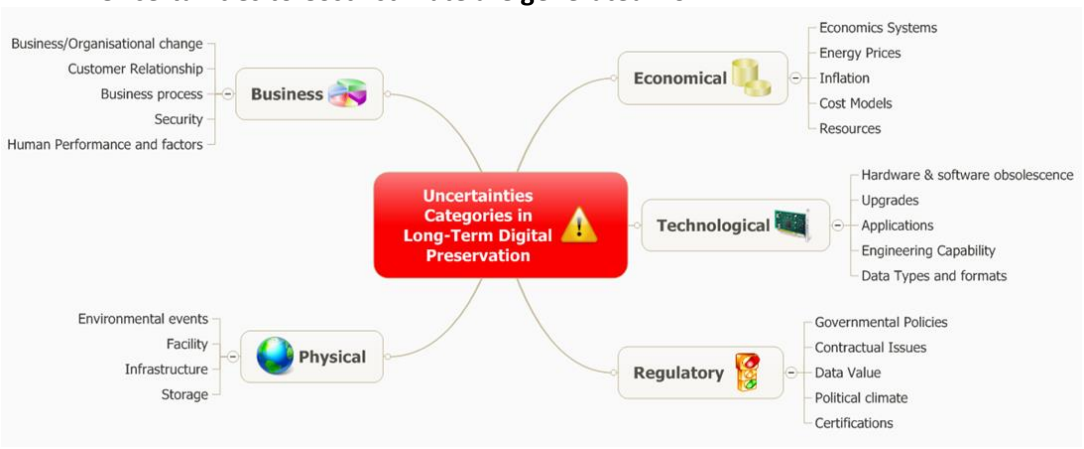
12. What are the main obsolescence issues that face your preservation process?

Obsolescence	Mitigation Strategy	Cost

13. What prioritisation category do you think is suitable to differentiate between these obsolescence issues? ……

A.4 Capturing Uncertainties in LTDP Cost Estimation

- **Personal Details**
  - ❖ Company: _____
  - ❖ Name/email: _____
  - ❖ Job Title: _____
  - ❖ Years of experience in Digital Preservation/IT: _____/_____
- **Uncertainties to Cost Estimate are generated from:**



Please identify the impact of uncertainties and their categories on the following Cost Breakdown Structures impact on cost

(***To provide impact use from 1 to 9. From 1 to 9; 1 represents that the obsolescence issue has least impact; 9 represents that the obsolescence issue has the highest impact***)

Cost Breakdown Structure Cost Element		Uncertainties Impact Uncertainty	Impact on cost
Initial Investment	Infrastructure Acquisition		
	Implementation		
	Redundancy		

Cost Breakdown Structure		Uncertainties Impact	
Cost Element		Uncertainty	Impact on cost
Infrastructure Acquisition Cost	Hardware (Network Equipment, Servers, Storage)		
	Software (OS)		
	Other (Power Distribution System, Cooling System, Power Failure Recovery, Fire Protection, security system and NOC)		

Cost Breakdown Structure		Uncertainties Impact	
Cost Element		Uncertainty	Impact on cost
Implementation Cost	Installation Cost		
	Configuration Cost		
	Commissioning Cost		
Redundancy Costs			

Cost Breakdown Structure		Uncertainties Impact	
Cost Element		Uncertainty	Impact on cost
Running Cost	Idle Power Consumption		
	Internet Connection		
	Maintenance Cost		

	Software Licenses		

Cost Breakdown Structure		Uncertainties Impact	
Cost Element		Uncertainty	Impact on cost
Preservation Activities Cost	Full Load Power Consumption		
	Staff Cost		

Cost Breakdown Structure		Uncertainties Impact	
Cost Element		Uncertainty	Impact on cost
Processing Cost	Public Cloud Providers		
	OS		
	Utilisation Cost		
	RAM		

Cost Breakdown Structure		Uncertainties Impact	
Cost Element		Uncertainty	Impact on cost
Ingest	Information Package Generation		
	Quality Check		
	Metadata Generation		

	Data Protection (Encryption, Anonymisation, etc.)		

Cost Breakdown Structure		Uncertainties Impact	
Cost Element		Uncertainty	Impact on cost
Storage	Data Transfer		
	Storage Volume		

Cost Breakdown Structure		Uncertainties Impact	
Cost Element		Uncertainty	Impact on cost
Data Management	Fixity Check		
	Reporting		
	File Deletion		
	Amendment to Metadata		
	Access Audits		

Cost Breakdown Structure		Uncertainties Impact	
Cost Element		Uncertainty	Impact on cost
Access	Information Package Retrieval		
	Data Protection		

Cost Breakdown Structure	Uncertainties Impact	
Cost Element	Uncertainty	Impact on cost

Transformation	Data Migration		
	Virtual Appliance Initiation		

A.5 Capturing Obsolescence in LTDP Cost Estimation

- **Personal Details**
  - ❖ Company: **Science and Technology Facilities Council**
  - ❖ Name/email:
    _____
  - ❖ Job Title: _____
  - ❖ Years of experience in Digital Preservation/IT: _____/_____

- **Defining Obsolescence in Long-Term Digital Preservation (LTDP):**

"Access to digital data becomes obsolete when it is supporting systems required for access, either hardware or software, can no longer be provided or supported by its original provider or a third party.
Also people skills and strategies can become obsolete when data cannot be retrieved or understood after preservation as expected due to long period of time and/or the evolution of technology".

1. Based on the Obsolescence definition, do you agree to the following breakdown of the main obsolescence issues in Long-Term Digital Preservation (LTDP)?
   ☐ YES ☐ NO
   - ❖ Main Obsolescence Issues:
     - Hardware
     - Software
     - People Skills
     - Preservation Strategies
   - ❖ If you have answered no please give the reason and other breakdown proposed

---

2. Please assess the impact of each obsolescence issue on total cost of digital preservation (***from 1 to 9; 1 represents that the obsolescence issue has least impact; 9 represents that the obsolescence issue has the highest impact***): please feel free to add extras at the bottom of the table

Obsolescence Issues	Impact on Cost								
Hardware	1	2	3	4	5	6	7	8	9
Software	1	2	3	4	5	6	7	8	9
Human Skills	1	2	3	4	5	6	7	8	9
Preservation Strategies	1	2	3	4	5	6	7	8	9
	1	2	3	4	5	6	7	8	9
	1	2	3	4	5	6	7	8	9

3. Please assess the impact of each hardware obsolescence issues on total cost of digital preservation (***from 1 to 9; 1 represents that the obsolescence issue has least impact; 9 represents that the obsolescence issue has the highest impact*** please feel free to add extras at the bottom of the table

Hardware Obsolescence Issues	**Impact on Cost**								
Whole System	1	2	3	4	5	6	7	8	9
Part of the system	1	2	3	4	5	6	7	8	9
Peripherals	1	2	3	4	5	6	7	8	9
Storage Media	1	2	3	4	5	6	7	8	9
	1	2	3	4	5	6	7	8	9
	1	2	3	4	5	6	7	8	9

4. Please assess the impact of each software obsolescence issues on total cost of digital preservation ***(from 1 to 9; 1 represents that the obsolescence issue has least impact; 9 represents that the obsolescence issue has the highest impact)***: please feel free to add extras at the bottom of the table

Software Obsolescence Issues	**Impact on Cost**								
File Formats	1	2	3	4	5	6	7	8	9
Applications	1	2	3	4	5	6	7	8	9
Plug-ins	1	2	3	4	5	6	7	8	9
Operating Systems	1	2	3	4	5	6	7	8	9
	1	2	3	4	5	6	7	8	9
	1	2	3	4	5	6	7	8	9

5. Please assess the impact of each human skills obsolescence issues on total cost of digital preservation ***(from 1 to 9; 1 represents that the obsolescence issue has least impact; 9 represents that the obsolescence issue has the highest impact)***: please feel free to add extras at the bottom of the table

Human Skills Obsolescence Issues	Impact on Cost								
Use of Software Applications	1	2	3	4	5	6	7	8	9
Use of Plug-ins	1	2	3	4	5	6	7	8	9
Extracting Information From Preserved Files	1	2	3	4	5	6	7	8	9
Operating Systems	1	2	3	4	5	6	7	8	9
	1	2	3	4	5	6	7	8	9
	1	2	3	4	5	6	7	8	9

6. Please assess the impact of each preservation plan obsolescence issues on total cost of digital preservation ***(from 1 to 9; 1 represents that the obsolescence issue has least impact; 9 represents that the obsolescence issue has the highest impact)***: please feel free to add extras at the bottom of the table

Preservation Plan Obsolescence Issues	Impact on Cost								
Hardware Systems Management	1	2	3	4	5	6	7	8	9
Software Applications/Systems Management	1	2	3	4	5	6	7	8	9
New Formats Management	1	2	3	4	5	6	7	8	9
Preserved Data/Information Selection	1	2	3	4	5	6	7	8	9
Transformation Monitoring and Actions Management	1	2	3	4	5	6	7	8	9
	1	2	3	4	5	6	7	8	9
	1	2	3	4	5	6	7	8	9

A.6 Second Delphi Workshop – Uncertainty & Obsolescence

Name: ……

Company: ……

Job Title: ……

Sector of Expertise:

☐ Clinical Trails ☐ Financial ☐ Healthcare ☐ Science Facility

1. What are the known costs endured with each activity

Activity Sub process detail	Cost …….	Cost …….	Cost …….	Cost …….	Cost …….
Data Management					
Pre-Ingest					
Ingest					
Access					
Active Preservation					

2. What are your current infrastructure resources available for the total lifecycle of preservation activities? Please mention any known:

Category	Cost …….	Cost …….	Cost …….	Cost …….	Cost …….
Physical space **(e.g. building hire, construction costs)**					
Costs of Hardware **(e.g. servers, racks, storage)**					
Costs of Software **(e.g. operating systems, database licencing)**					
Costs of Security **(e.g. physical or software)**					
Utilities requirements **(e.g. power, cooling)**					
Costs of Staff **(e.g. manager, technician, technology watch)**					
Other costs					

Uncertainties:

Uncertainties in cost estimation for LTDP can be defined as "*unexpected occurrence and/or the lack of enough knowledge about the processes carried-out, where this will generate noise in the cost estimated*".

3. What are the main uncertainties that face each preservation activity?

Uncertainty	Mitigation Strategy	Cost	Cost	Cost	Cost	Cost	Cost .....

4. What prioritisation category do you think is suitable to differentiate between these uncertainties? ……

Obsolescence

5. What are the main obsolescence issues that face your preservation process?

Obsolescence	Mitigation Strategy	Cost	Cost	Cost	Cost	Cost	Cost .....

6. What prioritisation category do you think is suitable to differentiate between these obsolescence issues?

……

A.7 Work breakdown structure Validation Questionnaire

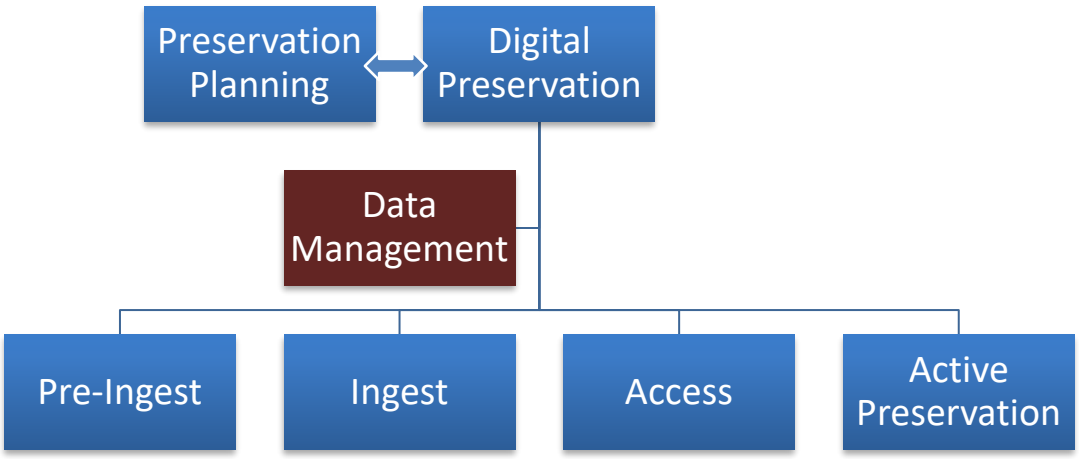
Names:

Companies:

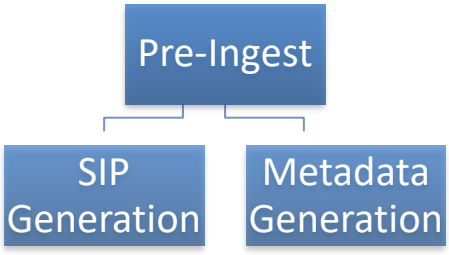
Sector of Expertise:

☐ Clinical Trails ☐ Financial ☐ Healthcare

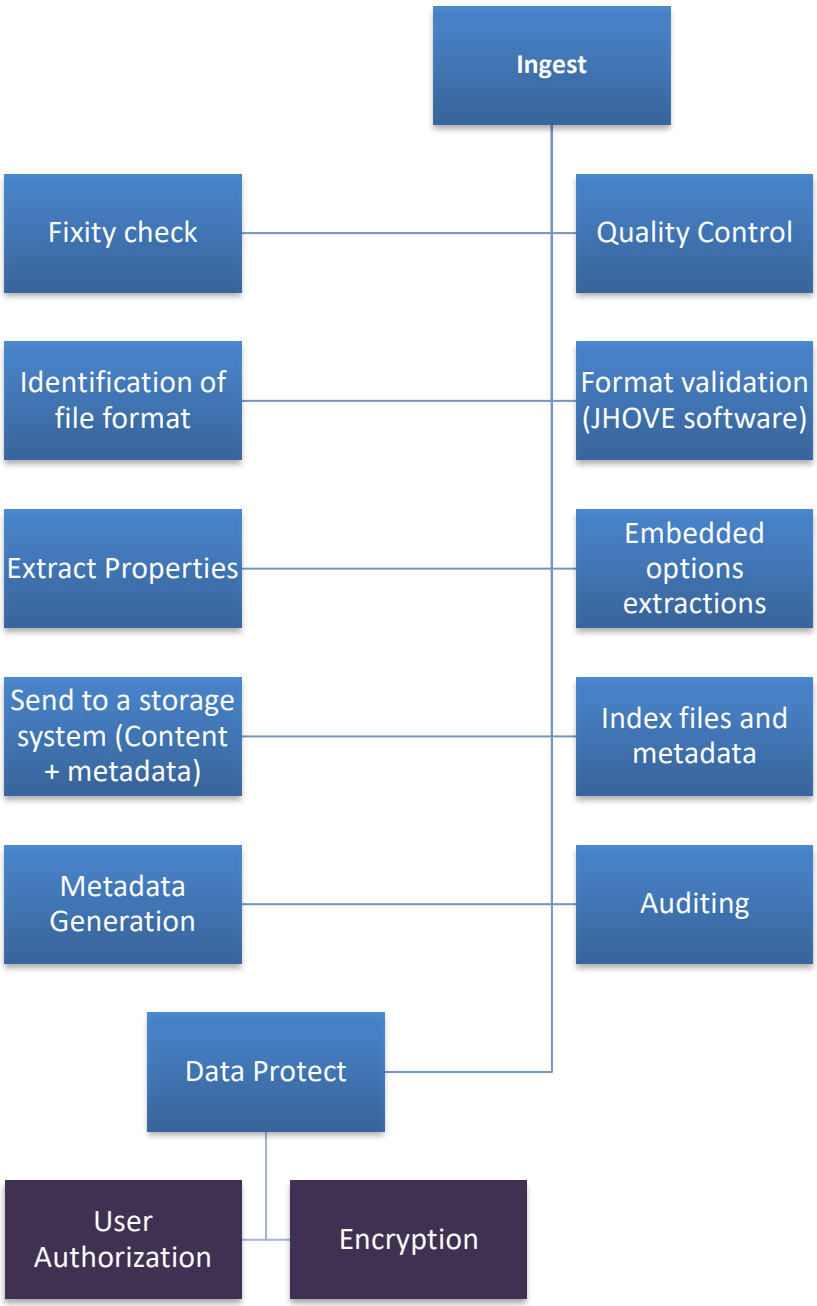
The following Figures 1 to 8 show the whole lifecycle Work Breakdown structure of Long-Term Digital Preservation for ENSURE



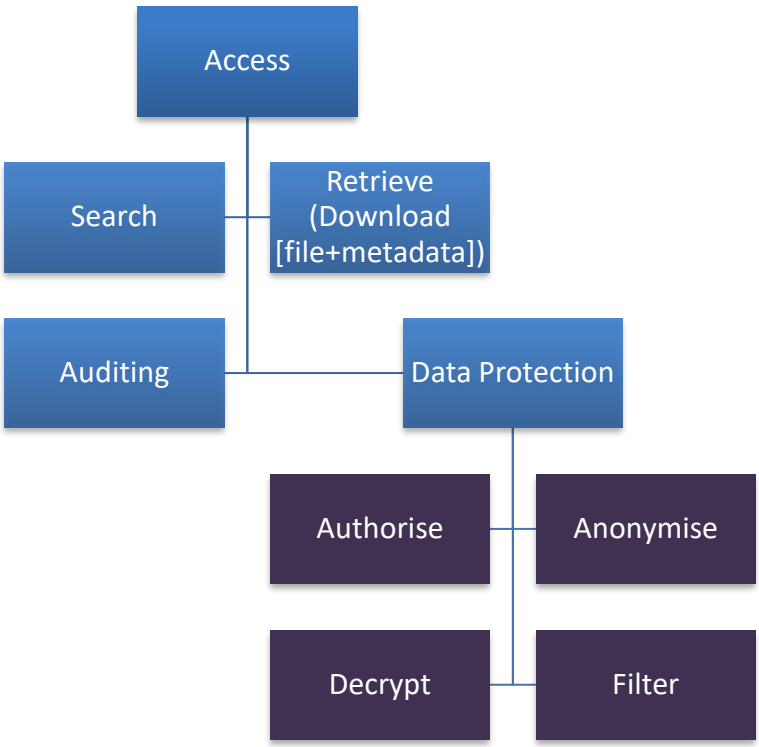
Pre-Ingest:



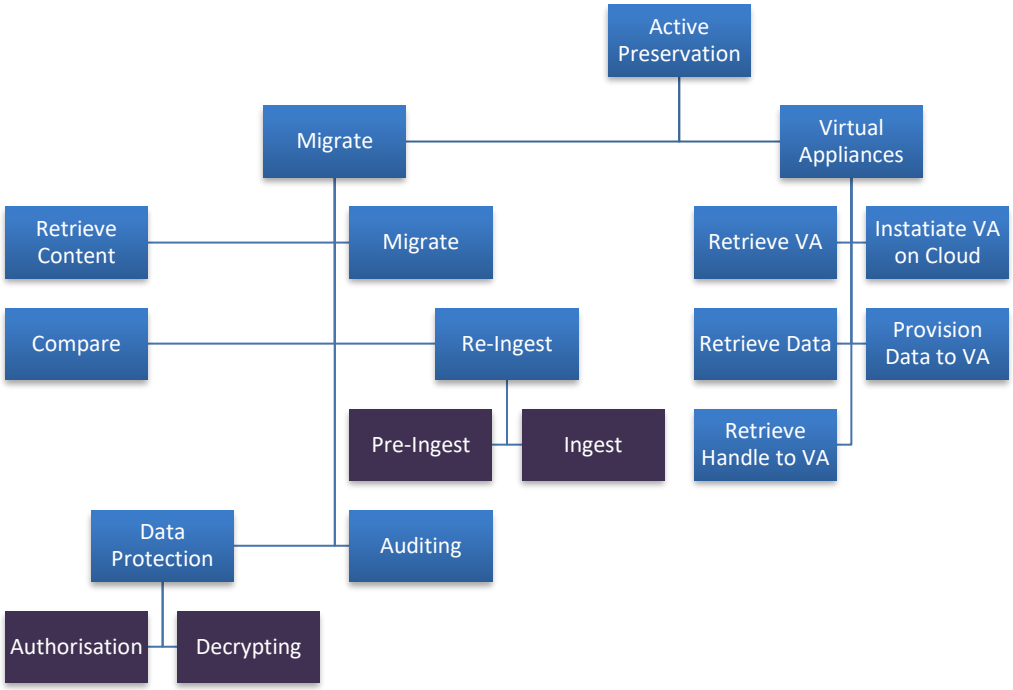
Ingest


Ingest
Fixity check
Quality Control
Identification of file format
Format validation (JHOVE software)
Extract Properties
Embedded options extractions
Send to a storage system (Content + metadata)
Index files and metadata
Metadata Generation
Auditing
Data Protect
User Authorization
Encryption

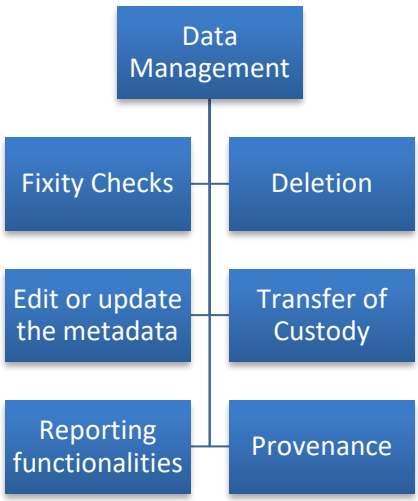
Access



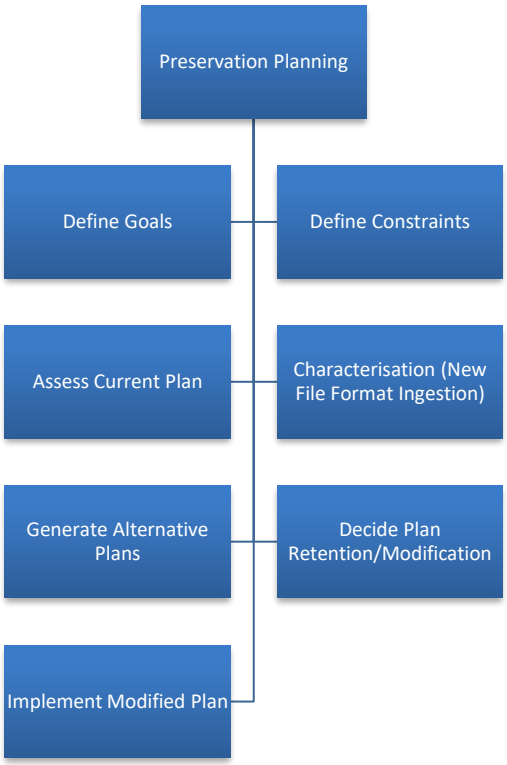
Active Preservation



Data Management



Preservation Planning/Configuration – Re-Configuration



A.8 Cost Breakdown Structure Validation Questionnaire

- **Personal Details**
  - ❖ Name/email (optional):
    _____
  - ❖ Company:_____
    _____
  - ❖ Job Title:_____
    __
  - ❖ Years of experience in Digital Preservation/IT:
    _____
    _____
    _____

- **Defining Private Cloud:**
  "*Private cloud is cloud infrastructure operated solely for a single organization, whether managed internally or by a third-party and hosted internally or externally*"
  _____
  _____

7. Based on your understanding of the functionality of a digital preservation system, do you agree to the following **Main Cost Elements** to a private cloud based preservation solution? If no please leave your comments

   ☐ Yes ☐ NO

   - ❖ Main Cost Elements of Digital Preservation on a Private Cloud:
     - Initial Investment Cost
     - Running Cost
     - Digital Preservation Actions

8. Based on your understanding of the functionality of a digital preservation system, do you agree to the following **Initial Investment** cost elements to a private cloud based preservation solution? If no please leave your comments

   ☐ Yes ☐ NO

   - Initial Investment Cost
     1. Infrastructure Acquisition Cost
        a. Hardware
        b. Software
        c. Other Costs
     2. Implementation Cost
        a. Installation Cost
        b. Configuration Cost
        c. Commissioning Cost
     3. Redundancy Requirements Cost
        a. Reduced Redundancy Cost

b. Back-ups Cost

9. Based on your understanding of the functionality of a digital preservation system, do you agree to the following **Running** cost elements to a private cloud based preservation solution? If no please leave your comments

☐ Yes ☐ NO

- Running Cost
    1. Power Consumption Cost
        a. Idle Power Consumption for Data Centre
        b. Cooling System Costs
    2. Internet Connection Cost
    3. Maintenance Cost
        a. Preventive
        b. Corrective
        c. Network Operations Centre Costs
    4. Software Licenses Cost

10. Based on your understanding of the functionality of a digital preservation system, do you agree to the following **Digital Preservation Actions** cost elements to a private cloud based preservation solution? If no please leave your comments

☐ Yes ☐ NO

- Digital Preservation Actions (*ingest, data management, transformation, access, etc.*)

    1. Action Algorithm Processing Time
    2. Full load Power Consumption of Data Centre
    3. Staff Utilisation Cost

End of Survey

Thank You for Your Kind Contribution

All information provided will be anonymous and personal information is protected under the data protection act 1998.

A.9 Framework Validation Questionnaire

1) Personal Details

a) Name: ……

b) Company: ……

c) Job Title/Role: ……

d) Years of experience in Digital Preservation/IT: ……

2) Case Study Description: (*Filled in meeting*)

a) Business Sector: ……

b) Initial / Annual Data Volumes: ……/……

c) Number of Back-ups: ……

d) Fixity Checks per Year: ……

e) Additional Information: ……

3) Framework Qualitative Validation:

a) Is the logic (process/rational) behind the Cost Estimation Framework valid?

1	2	3	4	5	6	7	8	9	10
Not Valid	Valid with major issues				Valid with minor issues				Valid

Comments:

……

b) Is the Cost Model suitable for your organisation's/business sector's digital preservation requirements?

1	2	3	4	5	6	7	8	9	10
Not Suitable	Suitable with major issues				Suitable with minor issues				Suitable

Comments:

……

c) Can the cost estimating framework be considered as generic to other business sectors that want to estimate the cost of long-term digital preservation? ☐Yes ☐No

d) Can the framework accommodate other business sectors? ☐Yes ☐No
If No then Why not……

e) Are the listed obsolescence issues in the framework complete? ☐Yes ☐No
If No then what is missing……

f) Are the listed uncertainty categories in the framework complete? ☐Yes ☐No
If No then what is missing……

g) Is the list of sector differences requirements acceptable? ☐Yes ☐No
If No then Why not……

h) What are the potential limitations and challenges for using the framework?

……

i) What are the strongest features of the framework?

……

j) What are the weakest features of the framework?

……

4) Tool Qualitative Validation

a) Does the validation tool provide sufficient initial information to validate the Framework's Concept? ☐Yes ☐No, If No why is it insufficient……

b) Are key cost drivers considered in the framework and the validation tool? ☐Yes ☐No If No then what is missing……

c) Does the tool reflect the adaptability required for it to be business sector independent?
☐Yes ☐No, If No then why not

……

d) Are the assumptions made in the tool acceptable? ☐Yes ☐No
If No then why not

……

e) Does the tool provide acceptable default values? ☐Yes ☐No
If No then what is missing……

f) Does the tool provide you with the ability to change these existing default values?
☐Yes ☐No, If No then why not

……

g) What are the potential limitations and challenges for using the validation tool?

……

……

h) What are the strongest features of the validation tool?

……

i) What are the weakest features of the validation tool?

……

5) Quantitative Validation

a) What's your evaluation of the output from the tool after populating it with information from your case study?

……

b) Is the cost estimate accurate enough for the purpose it was required for?

1	2	3	4	5	6	7	8	9	10
Not Accurate	Low Accuracy				Quite Accurate				Very Accurate

Comments:

……

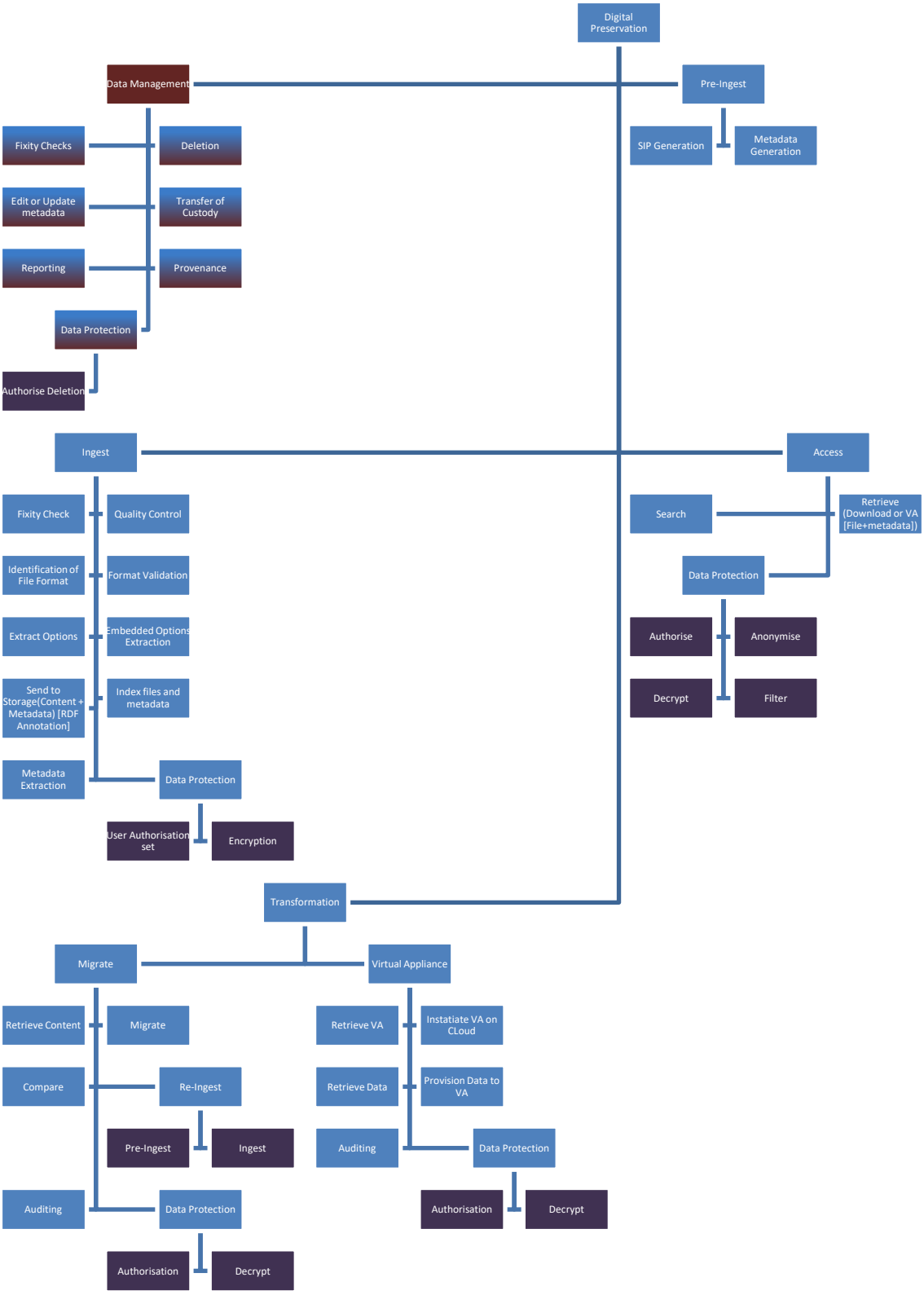
c) Is the framework accurate enough for the purpose it was developed for? ☐ Yes ☐ No
If No then why not

……

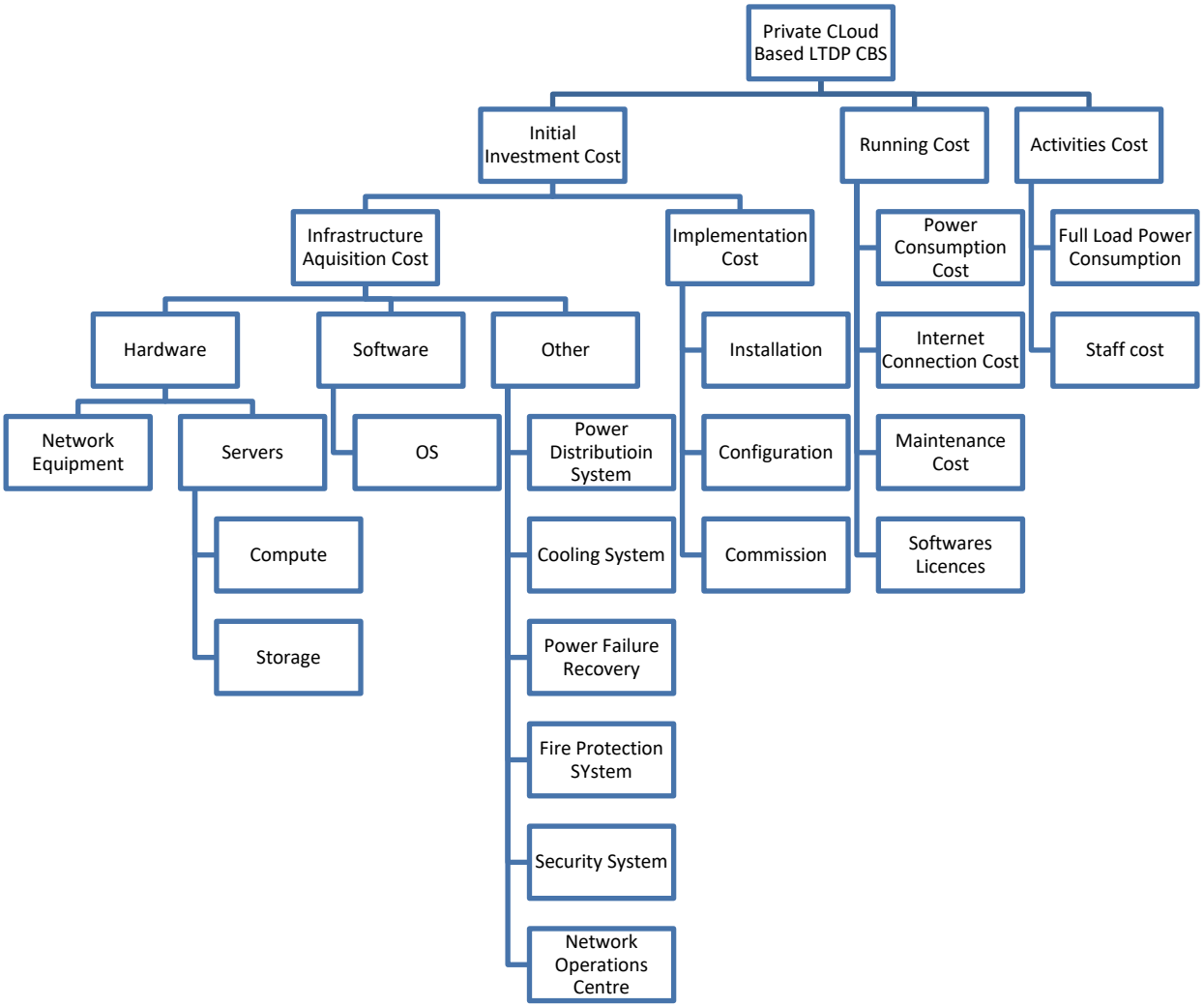
d) Please give anymore feedback/comments that you think are important:

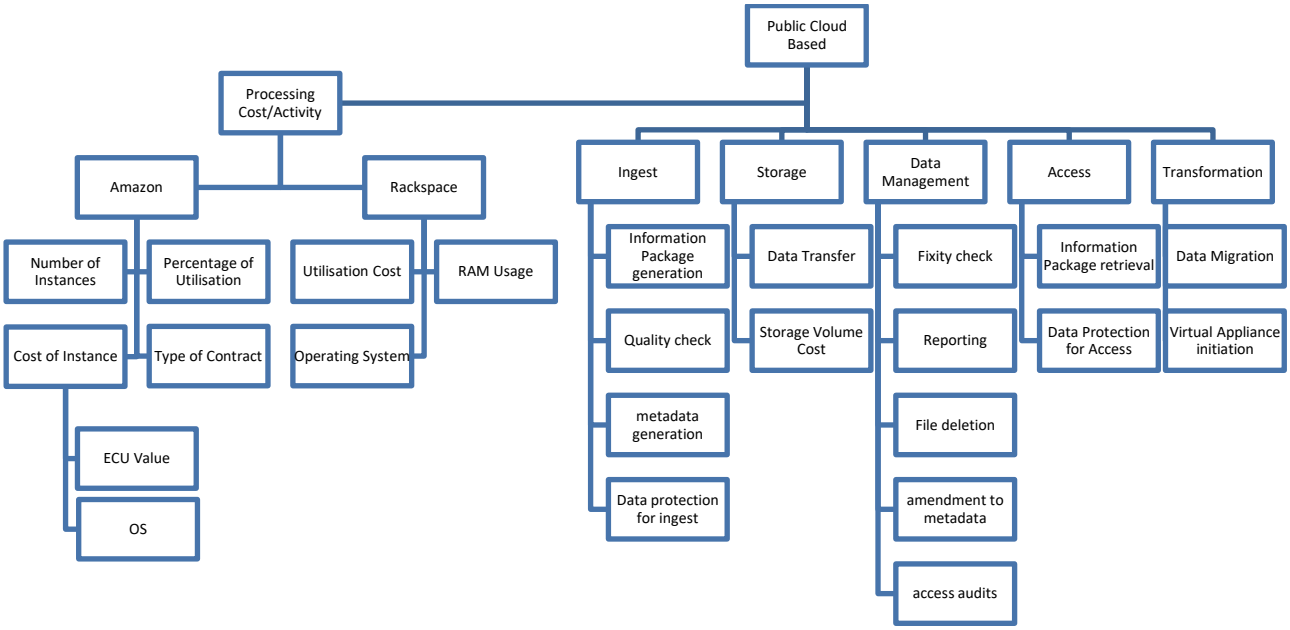
……

Appendix B Complete Work Breakdown Structure



Appendix C Complete Cost Breakdown Structure




Public Cloud Based
Processing Cost/Activity
Amazon
Rackspace
Number of Instances
Percentage of Utilisation
Cost of Instance
Type of Contract
ECU Value
OS
Utilisation Cost
RAM Usage
Operating System
Ingest
Storage
Data Management
Access
Transformation
Information Package generation
Quality check
metadata generation
Data protection for ingest
Data Transfer
Storage Volume Cost
Fixity check
Reporting
File deletion
amendment to metadata
access audits
Information Package retrieval
Data Protection for Access
Data Migration
Virtual Appliance initiation

Appendix D Related Terminologies (Ruusalepp, 2003)

- Access: Terms and conditions of granting permission to use data resources and collections in an archive. Access may be restricted in some instances because of copyright, confidentiality or statutory requirements.
- Acquisition: The official statement issued by an archive identifying types of data resources it will collect or acquire and the terms and conditions under which it will do so.
- Archive: An organisation that intends to preserve information for access and use by a designated community.
- Description: The process of capturing, analysing, organising and recording information that serves to identify, manage, locate and explain data resources and the contexts that produced them.
- Digital preservation: is a combination of techniques to actively management data to ensure its usability and accessibility over time (McLeod, R. et al. 2006). This means that it's not only the storage of data, but also making sure that this data is still meaningful as long as the data is needed to be kept.
- Documentation: All the material that provides information and guidance on how to interpret the contents of a digital resource, and which describes its contents, provenance, structure and other attributes.
- File Format: The specification of how the bits stored in a file should be interpreted.
- Life-cycle Concept: A concept that draws an analogy between the life of a biological organism, which is born, lives and dies, and that of a digital resource, which is created, used and then disposed of or transferred to an archives.
- Long-Term: A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and file formats, and of a changing user community, on the information being held by an archive.
- Metadata: "Data about other data." Metadata is collected with the purpose of explaining the technical and administrative processes used to create,

manipulate, use and preserve the data resource. It is often subdivided into further categories.

- Migration: The conversion of data resource from one hardware/software configuration or generation to another.
- Open Archival Information System (OAIS): An archive, consisting of an organisation of people and systems that has accepted the responsibility to preserve information and make it available for a Designated Community. The term ‘open’ in OAIS is used to imply that its reference model and standards are developed in open forums, and it does not imply that access to the archive is unrestricted.
- Operating Environment: All the hardware and software that is needed to run a digital resource.
- Preservation Copy: A copy made and used to preserve the intellectual content of a digital resource.
- Preservation Format: A format chosen for preservation purposes.
- Preservation Metadata: Preservation metadata – both technical and administrative – is kept to document the preservation processing in an archive, make it transparent and accountable and also in the preservation processing itself.
- Preservation Strategy: Coherent set of objectives and methods for maintaining digital components and related information over time, and for reproducing the related authentic data resources.
- Refresh: The process of copying digital resources from one storage medium to the same storage medium.
- User: Any member of the public who is allowed access to the archive and its holdings.
- Validation: Quality assurance performed by checking the contents of a file or a digital resource at the time of deposit or creation of preservation and/or dissemination versions

Appendix E LTDP Related Standards

E.1 **PREMIS**

This comprises a working group of over 30 members globally; their task has been to "*Define an implementable set of preservation metadata elements, with applicability in the digital preservation community*" (Riley, 2007). They built their work on OAIS. In May 2005, their final report was published and a data dictionary for the preservation of metadata was included (Riley, 2007). PREMIS is maintained by the US Library of Congress and consists of three parts (Higgins, 2009): the PREMIS data model, the PREMIS data dictionary and the PREMIS schema.

The PREMIS data model contains five entities. The model defines how they are related to each other.

1. Intellectual Entity: the digital object or parts of it.
2. Objects: a discrete digital information unit, a bit stream, file or representation (a file set to render an intellectual entity).
3. Events: an audit trail concerning changes made to a digital object throughout its lifecycle.
4. Agents: Persons, organizations or software that is responsible for preservation.
5. Rights: Permissions for the digital objects and their agents.

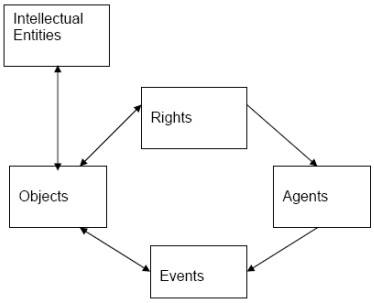


Figure - PREMIS Data Model (Riley, 2007)

The PREMIS Data Dictionary defines semantic units and semantic components in order to describe all of the entities. Finally, the PREMIS schema allows entities and their semantic units to be expressed consistently in XML.

The main focus of the PREMIS standard is metadata design. It is clear from the discussion that PREMIS complements the OAIS model and tries to improve the stability of the metadata design for the preservation system.

E.2 **NESTOR**

The aim of NESTOR is to introduce stable criteria for long-term digital repositories and to maintain these criteria over a long period. 14 catalogue criteria were generated at an abstract level (DCC 2010a). The basic principles in applying criteria (TRC 2006) are documentation, transparency, adequacy and measurability.

All objectives, specifications and implementations of the digital long-term repository should be documented, thus allowing evaluation.

Transparency is achieved by publishing parts of the documentation to allow users to check the degree of trustworthiness. Evaluation must be measurable; it is based on objectives and tasks to show its adequacy and trustworthiness.

NESTOR looks at ways to increase, measure and maintain the trustworthiness of repositories as an integral part of any preservation system.

E.3 **DRAMBORA**

This is a web toolkit to help audit repositories through a self-assessment process (DCC 2010b) by defining the scope of functions, identifying activities and assets, identifying the incorporated risks and vulnerabilities in the activities and assets, assessing and calculating risks, defining risk management measures and producing a report.

DRAMBORA's principal application (DCC 2010b) is to validate the effectiveness of repository infrastructure, to help plan for improvements, prepare for external audits and anticipate in planning development.

E.4 **TRAC**

TRAC provides tools for the audit, assessment and potential certification of digital repositories (DCC 2010c). TRAC provides tools for auditing and certifying digital repositories, establishes documentation requirements, has precise certification indicators and establishes suitable methodologies for determining the strength of digital repositories (DCC 2010c).