TOWARDS A COST MODEL FOR LONG TERM DIGITAL PRESERVATION

Mohamed Badawy [(1)]
Essam Shehab [(1)]
Paul Baguley [(1)]
Michael Wilson [(2)]

m.badawy@cranfield.ac.uk
e.shehab@cranfield.ac.uk
p.baguley@cranfield.ac.uk
michael.wilson@stfc.ac.uk

[(1)] Manufacturing & Materials Department
Cranfield University
Cranfield, Bedfordshire
MK43 0AL, UK

[(2)] Science & Technology Facilities Council
Harwell Oxford
Didcot
OX11 0QX

ABSTRACT

Digital preservation is a research area attracting interest due to its importance to a lot of business sectors. Interest is increasingly coming from firms and institutes looking to preserve their digital data for a long period of time, thus increasing the importance of accurate estimation of costs for carrying out preservation activities. A reason for this interest is the spiralling amounts of digital data available at companies, firms and organisations with commercial value, where preserving this ever growing digital population is becoming a major problem.

Estimating the costs for long-term digital preservation will enable decision makers to choose carefully what data to invest in preserving, for how long and what type of preservation techniques are best applied for their information. To address this need, a cost model is being developed to estimate costs for long-term digital preservation activities using storage in the cloud and taking into consideration the impact of mitigating uncertainties, especially obsolescence, issues on future costs. This cost model is part of the European project titled “Enabling kNowledge Sustainability Usability and Recovery for Economic value” (ENSURE) whichaims to providing a total long-term digital preservation solution for companies and public sector organizations interested in keeping their information alive for long periods of time.

This paper presents an overview of the work done so far to build the cost model, starting off with a quick comparison of other existing cost models, then a discussion about how the development of this cost model was approached; passing through the current state of the model and ending by summarising future work to be done to the model.

Key words: Long-Term Digital Preservation, Cost Model, Cost Estimation, Digital Curation, Cloud

1 INTRODUCTION

Long-term information preservation has been a normal activity for humans throughout time, but nowadays the amount of information produced has increased dramatically, because of advances in communication technologies and science. A large amount of generated data can be considered digital, since either it is originated digitally or it was digitized.

Any information produced will have had some costs incurred by the producer. This cost will increase if information is lost and had to be reproduced again, or if the data cannot be disclosed as evidence, when required, in court. This situation has forced many organizations to research investing in preservation, thus

the study of future costs of long-term digital preservation could have very broad impact. The importance of the cost study for an organization is determined by the importance of the information to be preserved and how long they are looking to preserve this data. Digital preservation's main goal is keeping digital information usable at some point in the future in spite of any obsolescence of hardware, software, processes, format and/or people skills. Long-term digital preservation is not an issue solely for information suppliers but an activity that involves many stakeholders.

The Open Archival Information System (OAIS), ISO 14721, is considered the most important standard for digital preservation (CCSDS, 2003). It was developed by the National Aeronautics and Space Administration (NASA)'s Consultative Committee for Space Data Systems (CCSDS). OAIS, Figure 1, defines a framework for a successful repository (Higgins, S. 2009). 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).

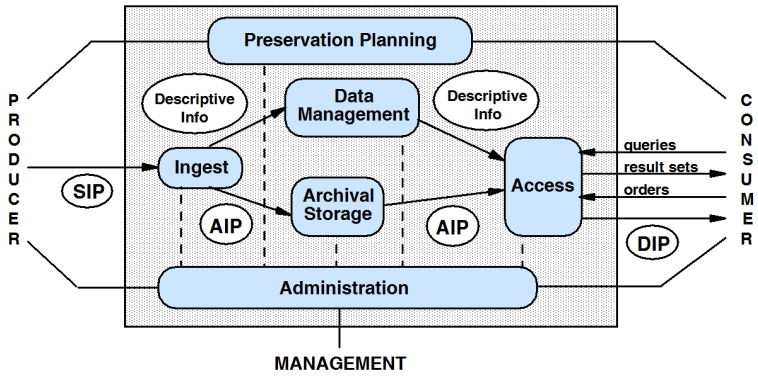

Figure 1: OAIS Functional Entities (CCSDS, 2002)

It is very important to know the activities and workflow of digital preservation. However predicting the costs of the long-term digital preservation is a complex and a difficult task. The cost model is expected to provide a reasoned case for digital preservation based on the cost elements identification (Xue, P et. al, 2011).

A cost model for digital preservation should be based on a well-defined breakdown of all activities covering both the costs of the preservation system and the costs of, for example, people's skills, management and administration. It must take into account and give the life cycle feedback to the companies on strategies and policies decision making, budget planning, and the change of the technology and activities. Seeking the real and true cost and being able to justify it is difficult and expensive (Russell and Weinberger, 2000).

Following in this paper a discussion about other related cost models, how is the ENSURE cost model being attempted and developed and finally a conclusion of work with a summary of future work.

2 COST MODELS FOR DIGITAL PRESERVATION

Information produced by different producers, will have production processes different to each other, thus might cost differently according to business sector, company or organization or project (Chapman, 2006). Many projects containing cost models to estimate costs of digital preservation exist for specific sectors, but most of them were targeting the libraries, national archives, representing the heritage sector and for laboratories and research facilities, representing science facilities sector.

These projects have main four cost models, NASA's Cost Estimation Tool (**CET**) (Hendley, 1998), Lifecycle Information for E-Literature (**LIFE**) (Wheatley, et. al, 2009), Cost Model for Digital Preservation (**CMDP**) (Kejser, et. al, 2011) and Keeping Research Data Safe (**KRDS**) (Stanger, 2011).

CET is the oldest and targeting scientific sector. It was developed to estimate lifecycle costs of maintaining scientific data centres. Its tool has a comparable database of historic data reachable through a set of what-if choices and parameter sensitive tests. The cost date comparable database that the CET has given a strong backing to the cost model, this database started with 29 projects data, and the model is constantly updating and adding new ones to them. 94 metadata fields are utilised to capture finest cost variations.

The main draw back that CET is suffering from is lack of an uncertainty study to make the model future proof, also CET is designed to serve space and earth observation so it's not flexible to accommodate the industrial or commercial sector (Hendley, 1998).

A highly recognised cost model that served the heritage sector is the LIFE project, which was developed by the cooperation of University of London and British library on three phases, LIFE^1, LIFE^2 and LIFE^3. Its main target sector to serve is libraries. The LIFE cost model looks at the complexity of the file format which it divides into 10 separate complexity levels. LIFE depends hugely on the OAIS standard model, which helped in breaking down the process levels.

The cost model is given as a spread-sheet, this is its main drawback since this interface is static and lacks the user interface, also the user's input is very limited which makes the model very similar to a financial report (Wheatley, et. al, 2009). LIFE^3 is being aimed to be more a more generic model.

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

The Danish National Archive has developed in 2011 a cost model, on two main stages, CMDP^1 and CMDP^2, CMDP^1 is for the Preservation Planning and Digital Migration and CMDP^2 is focusing on the ingest phase (CMDP, 2010). The model is based on the OAIS standard model and uses the activity based modelling technique (Kejser, et. al, 2011).

The previous cost models have three weakness in common primarily they didn't have any uncertainties integration in the model, which renders the model not future proof. The second weakness is that they did not take in consideration cloud storage as a storage option, which for the time being is the future of storage because of its ease of access and fast set-up time. Finally the third weakness is the lack of a cost techniques comparison, which should give a deeper insight towards the best estimation technique available.

3 ENSURE COST MODEL

The main differences between the ENSURE cost model and other cost models available are:

[1] Targeting storage on the Cloud.
[2] Investigating different cost estimation techniques
[3] Generic across more than one business sector
[4] Incorporates uncertainties and especially obsolescence issues mitigation costs

3.1 Research Methodology

To achieve the requirements and targets of the model a research methodology was put in place, to achieve the maximum out of the time dedicated for this project. Figure 2 shows how the main areas of the methodology are interacting together.

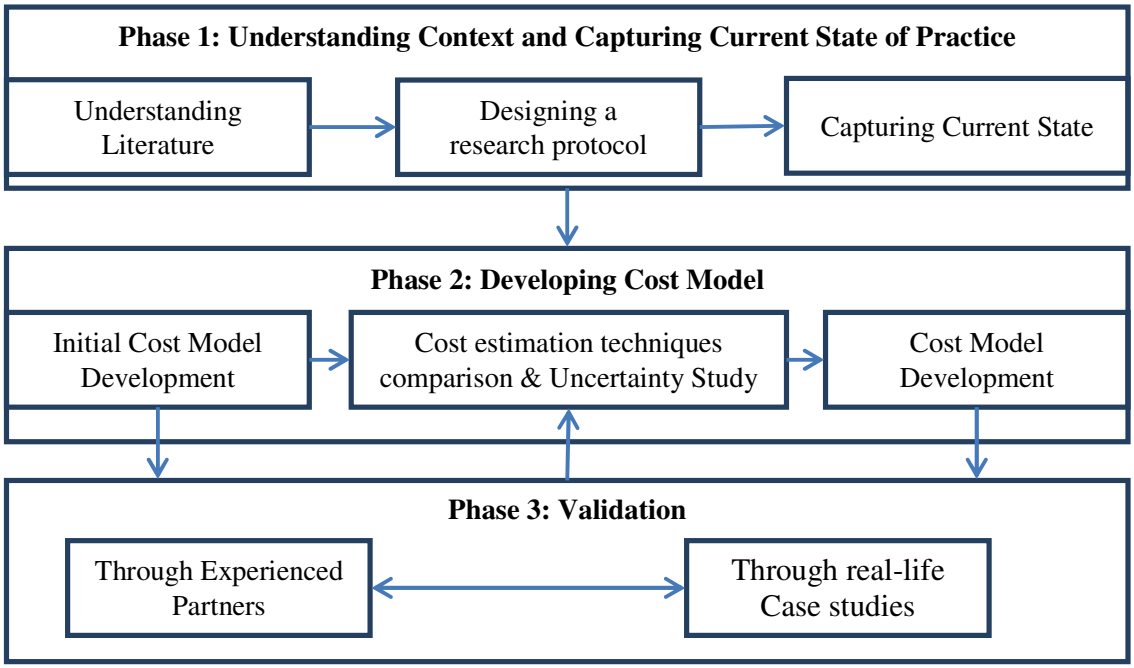

Figure 2: ENSURE Cost Model's Research Methodology

Three main phases construct the total research methodology, phase one focuses on understanding state-of-the art of the science of digital preservation cost modelling, and then capturing the current state of practice in the industry. The AS-IS model will be captured using a surveying technique, in this case a questionnaire. This questionnaire will be piloted first with some of the project partners with the most significant preservation experience.

Phase two will follow starting with developing an initial cost model that can model a simple form of digital preservation as a starting point. No uncertainties or obsolescence issues are integrated at this stage. The initial cost model should reflect straight forward and simple digital preservation processes. The work and cost breakdown structures and a comparison between cost estimation techniques should be implemented. This cost model should be validated with the experienced partners.

Finally in phase three validating the developed cost model will be through experienced project partners and real-life case studies.

3.2 Work Break-down Structure (WBS)

Piloting the research protocol resulted in a WBS which is shown Figure 3. The WBS resulted in showing five main stages of preservation activities to be carried out.

- Digital Preservation:
  - Data Management
  - Pre-Ingest
  - Ingest
  - Access
  - Active Preservation (Limited now for Migration)

The following flow chart in Figure 3 shows that data management activities are always active, while all other main activities must be called upon. Data management is an integral part of maintaining the preserved information that's why it is designed to be always active and carrying out checks, sorting, editing and reporting.

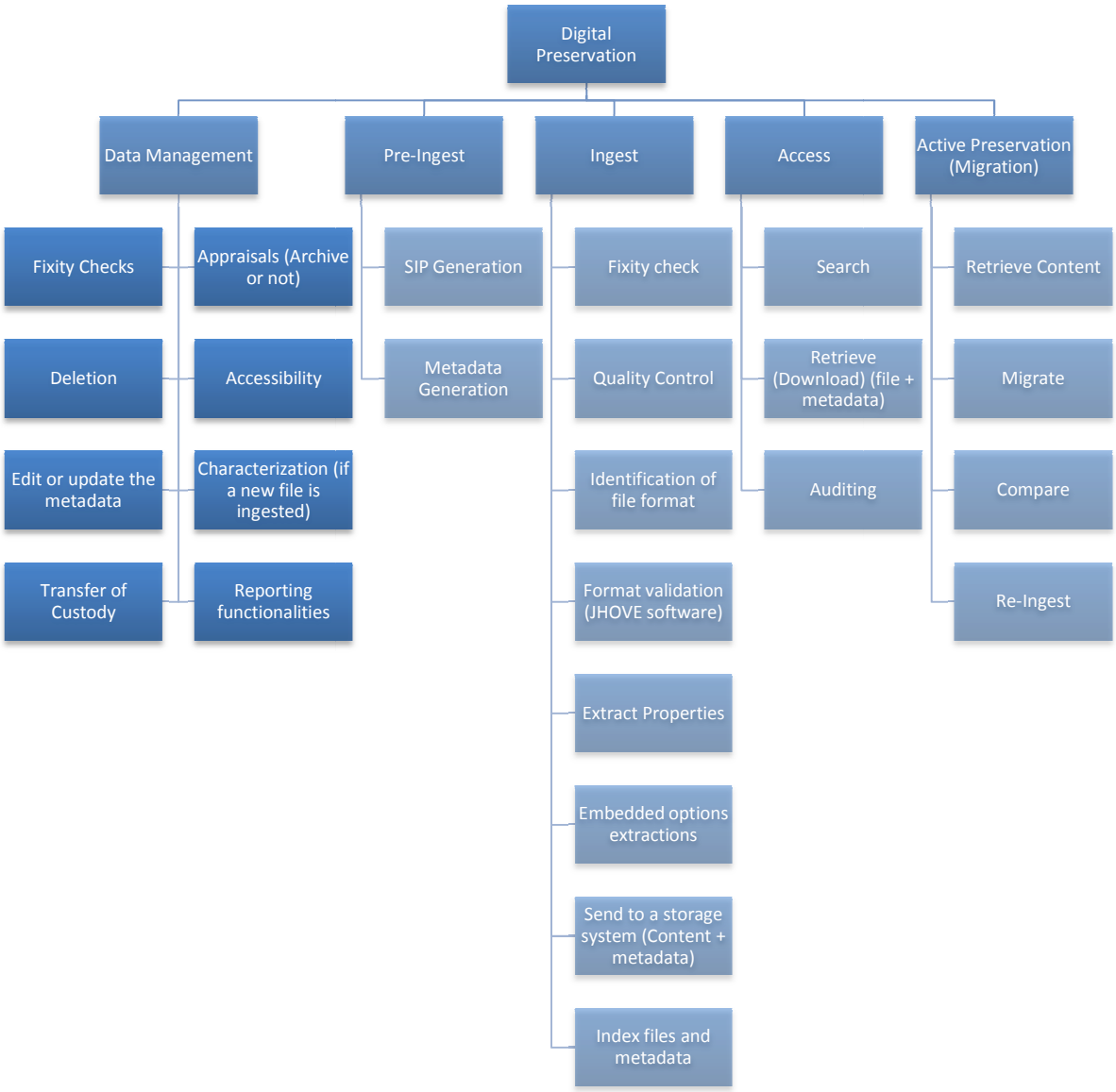

Figure 3: Long-Term Digital Preservation Work Break Down Structure

3.3 Initial Obsolescence Study

After piloting the research protocol with the experienced partners, a short list of obsolescence issues was generated. The main areas of obsolescence are shown in Figure 4.

- Software
  - File Formats
  - Applications
  - Plug-ins
  - Operating System
- Hardware
  - Whole System
  - Part of the system
  - Peripherals
  - Storage Media
- Human Skills

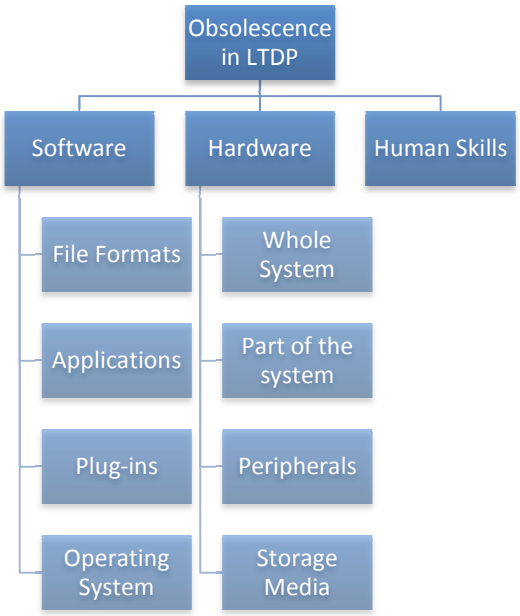

Figure 4: Long-Term Digital Preservation Work Break Down Structure

4 ENSURE INITIAL COST MODEL

The initial cost model is fairly simple and straight forward formulae, based on some activities that have been spotted within the WBS. This initial cost model takes public cloud storage facilities as its only storage option.

It is expected that the total preservation cost is going to be represented by the cost of ingesting, the cost of archiving and the cost of accessing the data.

Total Preservation Cost = Ingest Cost + Archival Cost + Access Cost + Data Management Cost (a)

The Ingestion cost is composed of the information packages generation cost, both Archival Information Packages (AIP) and Submission Information Packages (SIP), information package quality assurance cost and the cost generation of description data.

Ingest cost = Information Packages Generation Cost + Quality Check Cost + Description Data Generation Cost (b)

While the archival cost can be expected to be storage costs, quality assurance costs, preservation actions costs, e.g. migration, and the cost of data transfer (especially for cloud).

Archival Cost = Data transfer cost + Storage Cost + Transformation Cost (c)

Data Management Cost = Fixity Check Cost + Reporting Cost + File Deletion Cost + Amendment to Metadata Cost + Auditing Access Costs (d)

Finally the access cost is the cost of retrieving the information package, the cost of delivery response and the cost co-ordinating access activities.

Access Cost = Information Package Retrieval Cost + Delivery Response Cost (e)

Where:

- **Data Transfer = Cost per GB* x Storage Volume in GB***
- **Storage Cost = Storage Duration (months) x Cost per month per GB* x Storage Volume**
- **Fixity Check Cost = Cost per Test x (Storage Duration (months) / (Frequency of Test (months)))**
- **Test Cost = Cost Per hour x Test Duration (hours)**
- **Test Duration (hours) = Storage Volume (GB*) / Test Processing Rate (GB/hour)**

**GB = Giga Bytes*

5 ENSURE COST MODEL CHALLENGES

To produce a cost model of this type for long-term digital preservation, many challenges have been met. A summary of these challenges are shown in Figure 5. These challenges are classified into four main categories, new businesses interested in digital preservation, lack of cost information, many cost estimation methodologies and technology based challenges.

Interest is rising by new business sectors, such as healthcare, manufacturing, financial and clinical trials sectors, due to the importance of their digital information and due to legislations that require them to keep this information for a minimum period.

Uncertainties is something that was never taken in consideration before when designing a cost model for digital preservation, this is because the ENSURE cost model is the only model being design for the long-term, thus facing uncertainties mitigation cost is an important issue. This is coupled by the difficulty to acquire cost information from commercial companies due to their eagerness to keep their private data hidden. Each sector can provide more than one file format and each will have a different cost.

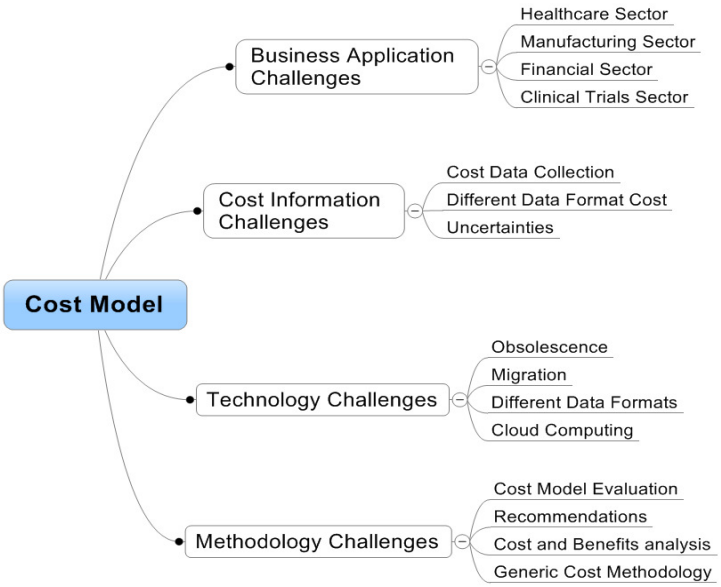

Figure 5: Challenges of ENSURE Cost Model for Long-Term Digital Preservation (Xue, et.al, 2011)

Technology challenges are the fastest changing challenges of them all, since it deals with market shifts. This is cause mainly by obsolescence of many components of the preservation system, software, hardware, organization processes and policies or human skills as seen in Figure 4. The other issue is the emerging cloud computing technology, which was never implemented in any digital preservation solution.

A new cost model that adopts new recommendations, cost/benefit analysis and trying to be generic will face some methodology challenges.

6 CONCLUSIONS AND FUTURE WORK

This paper has discussed the initial cost model that is part of the European project ENSURE. ENSURE aims at providing a total solution for the long-term digital preservation needs for companies and organizations in the healthcare, clinical trials and financial sectors. It adopts state-of-the art technology by incorporating cloud computing solutions. Since ENSURE is trying to provide a total solution it contains a cost model which feeds into a cost benefit analysis component to provide decision makers with the information they need to evaluate the options for digital preservation. The cost model component must be able to meet those challenges faced by the overall project, such as the new cloud computing technology, and potential obsolescence issues. This breadth might make the model generic enough, not only to meet the requirements of those market sectors that the current project is addressing, but also to have the ability to apply to wider industrial sectors, such as manufacturing.

Two pieces of future research work are planned to improve the cost model. Firstly, a thorough investigation of the available cost estimation techniques, which will improve the accuracy and flexibility of the cost model. Secondly, to expand on the initial obsolescence study in order to add to it a deep uncertainty analysis, in order to help elevate the cost model from providing only a single point estimate to providing three points estimate.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 270000 to the ENSURE project which involves 13 research and industrial partners (www.ensure-fp7.eu).

REFERENCES

- 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 18/01/2012)
- Chapman, S. (2006). Counting the costs of digital preservation: Is repository storage affordable? Journal of Digital Information, 4(2) p1-15
- CMDP Official Website, 2010, http://www.costmodelfordigitalpreservation.dk/ (Last visited 26/01/2012)
- Hendley, T. (1998). Comparison of methods and costs of digital preservation. A JISC/NPO Study within the Electronic Libraries (eLib) Programme on the Preservation of Electronic Materials
- Higgins, S. (2009), "PREMIS Data Dictionary for Preservation Metadata"
- Kejser, U. B., Nielsen, A. B., & Thirifays, A. (2011). Cost model for digital preservation: Cost of digital migration. International Journal of Digital Curation, 6(1), p 255-267
- Russell, K., & Weinberger, E. (2000). Cost elements of digital preservation (Online Draft), http://www.scribd.com/doc/7345161/RUSSELL-Kelly-Cost-elements-of-digital-preservation, (last visited 09/01/2012)
- Wheatley, P. and Hole, B. (2009). LIFE3: Predicting long-term digital preservation costs. http://www.life.ac.uk/3/docs/ipres2009v24.pdf (last visited 17/01/2012)
- 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