EVALUATION OF THE COST OF PRESERVATION PROJECT

Final Report - December 2009

Prepared by:

Neil Beagrie, Daphne Charles and Najla Rettberg

Charles Beagrie Limited

www.beagrie.com

Contents

1. Executive Summary

1.1. Introduction

This evaluation report reviews the outcomes of the Cost of Digital Preservation project. The aim of this report is to carry out an impartial appraisal of the project. Consultants from Charles Beagrie Limited conducted this evaluation between November and December 2009.

The report looks at project goals and to what extent they were achieved and has carried out an additional independent benchmarking exercise against related costing approaches (LIFE, KRDS, NASA CET). Finally future opportunities and areas of synergy with similar projects have also been considered and recommendations made.

1.2. Key Findings and Recommendations

1. The project report is a welcome addition to the literature on digital preservation costs and cost modelling. This area of study is widely seen to be important for digital collections in ALM institutions but very few significant projects have been undertaken to date.

2. Work completed to date provides a valuable foundation for future work to build on and leverage this initial project. We have made a number of suggestions which we hope will be helpful in any future continuation.

3. The project has successfully completed its success criteria for analysis of the LIFE project and related costs models. However it was found necessary to adjust the other two key success criteria during the course of the project. This was due to a number of factors including less help than expected in existing models, budget stops, and staff turnover. In addition, with hindsight, the project's initial objectives may have been too ambitious for the resources (60,000 euros/ 9 person months) and timescale (12 months) allocated.

4. We have made a number of suggestions and recommendations for future work in section 5 in the following areas:

    - Student study placements to process or create additional cost data sets for Danish ALM institutions;

    - Refining the accuracy of the prediction model;

- Refining Format Complexity;
- Extending work to include Cost/benefit Analysis;
- Developing dual applications of the CMDP: a high-level cost management application and a detailed process application;
- Exploring Scale issues and related costs and impacts on areas such as workflow and bit preservation;
- "Digital Continuity": the maintenance and sustainability of long-lived current digital information in government and ALM institutions. The potential applications of cost models to this type of activity might be a future application for CMDP;
- Collaboration with other Projects. We note the potential for various collaborations with other projects. This might be extended to funded partnerships where funding conditions allow, as well as mutual in-kind exchange of experience.

We believe the CMDP has produced interesting results with relatively modest resources and should be continued to build on this initial work.

2. Introduction

This study reviews the outcomes of the Cost of Digital Preservation project carried out by The Royal Library, (KB), the State and University Library (SB) and the Danish National Archives (DNA) for the development of a generic cost model for the preservation of digital materials at archives, libraries and museums (ALM) institutions. The project ran from November 2008 to December 2009 staffed by project officers from DNA and KB. A need was expressed by the institutions for a more comprehensive understanding of the costs of digital preservation in order to increase efficiency in managing collections. The project developed an independent model, called the Cost Model for Digital Preservation, (CMDP) and looked in particular at the OAIS standard (CCSDS 2002) to identify price critical functions and to put them into operation. The model was tested on two separate migration cost data case studies.

A request for an external project review was put forward by the project team. Thus the aim of this report is to carry out an impartial appraisal of the project. Consultants from Charles Beagrie Limited conducted this evaluation between November and December 2009.

The first section of the report looks at project goals and to what extent they were achieved. This is in two parts, a self-evaluation by the CMDP project team (prepared from statements in the final project report), and an evaluation and commentary on the final report by the external evaluators.

While the CMDP project looked in detail at other cost approaches such as Keeping Research Data Safe (KRDS), LIFE, and the NASA Cost Estimation Took (NASA CET) and how they could be integrated into the study, this evaluation process has carried out an additional independent benchmarking exercise against these related approaches.

Finally, future opportunities and areas of synergy with similar projects have also been considered and recommendations made.

3. Project Self-Evaluation against Original Objectives

Appendix A provides a self-evaluation against original objectives by the CMDP team. To complete Appendix A, we have assessed available project documentation and pulled together from this:

- the original goal, success criteria, and supporting goals and activities of the project; and
- the CMDP team's own evaluations of how well or completely they were realised.

The following project documents were analysed to achieve this:

- Final project report (Kejser el al 2009a);
- Spreadsheet of equations and cost dependencies(Kejser el al 2009b);
- Conference article, 'Cost Model for Digital Curation: Cost of Digital Migration'. (Kejser el al 2009c);
- Power point presentation(Kejser el al 2009d);
- Analysis of OAIS paper (Kejser el al 2009e).

The project report highlights in its executive summary that the project had one main goal:

> "..to set up a generic model for the calculation of costs related to preservation of digital materials at ALM institutions." (Kejser et. al, 2009a, p 2).

It then lists three success criteria for fulfilment (Kejser et. al, 2009a, p 4):

- A closer investigation of the project LIFE and its follow-up LIFE 2,as well as related cost models;
- Development of a cost model which is tested in detail on Danish data and experiences;
- Use of the cost model on the estimated amounts of digital material which must be preserved for the state archives, libraries and museums in the next ten years.

A number of ‘Goals for the cost model’ were also listed (Kejser et. al, 2009a, p 7), as well as a series of ‘Major activities’ within the project (Kejser et. al, 2009a, p 3). This self-evaluation has considered all these as sub-goals within the project. In total, twenty-one original objectives (goal, success criteria, and sub-goals) for the project were identified and tabulated below.

Note that the while the goals are quoted exactly as written in the report, small amounts of the text have been edited to make the table in Appendix A more concise.

4. External Consultants Evaluation and Commentary

4.1. General Comments

1. The project report is a welcome addition to the literature on digital preservation costs and cost modelling. This area of study is widely seen to be important for digital collections in ALM institutions but very few significant projects have been undertaken to date.

2. Work completed to date provides a valuable foundation for future work to build on and leverage this initial project. We have made a number of suggestions which we hope will be helpful in any future continuation.

3. The project has successfully completed its success criteria for analysis of the LIFE project and related costs models. However it was found necessary to adjust the other two key success criteria during the course of the project. This was due to a number of factors including less help than expected in existing models, budget stops, and staff turnover. In addition, with hindsight, the project's initial objectives may have been too ambitious for the resources (60,000 euros/ 9 person months) and timescale (12 months) allocated.

4. An additional factor that should be considered by the project is that related costs projects may not have given sufficient visibility to the resources and timescale needed either to develop their models or to capture and process cost data for analysis, leading to some under-estimation for these tasks in CMDP project. We recommend that information on resources and timescales are gathered from related projects on resources and timescales to develop cost models and/or process test cost data to help inform any second phase of CMDP. This will supplement and extend lessons learned in CMDP itself.

5. The section numbering and sub-headings of the report sometimes reach down to four numbered levels and a fifth un-numbered level. It can be confusing and difficult to follow the structure of the document particularly in section 3.2. We recommend headings and section numbering are reviewed and streamlined in the final version.

4.2. Specific Comments

Section 2 .1 Life Costing Model

6. Existing text and discussion of the LIFE model is well considered and accurate. Consideration could be given to a short discussion of the plans for LIFE 3 and more discussion of the perceived strengths and weaknesses overall for CMDP in section 2.1.3 and reasons for a separate development. Resources and timescales for LIFE1, LIFE2, and LIFE3 might also be obtained and added here.

Section 2.2 Keeping Research Data Safe (KRDS1)

7. Existing text and discussion of the KRDS1 model is well considered and accurate. Consideration should be given to a short discussion of objectives for KRDS2 and any lessons for a future CMDP project. Resources and timescales for KRDS1 and KRDS2 might also be added here (5 months £64,000 excluding VAT; and 9 months and £52,000 excluding VAT, respectively).

Section 2.3 NASA Cost Estimation Tool

8. Existing text and discussion is accurate. The fact that NASA CET has calculated its error margin is very important (compared to LIFE which has been criticised for producing figures with apparent accuracy). It is also important that NASA CET has refined the model over a period of years and it is believed that the current error margin cannot be improved and to note the fact that the level of accuracy in other costs models is also likely to have a limit. Other statistical observations in NASA CET are relevant to CMDP such as the number of cost datasets needed to generate the model. A perceived drawback to the NASA CET approach is the need to update these cost datasets as further automation amends them over time. These issues may also be a factor for CMDP to consider for future stages of work. Resources and timescales for NASA CET might also be added here (information from NASA - The level of investment is roughly 2FTEs per year/$250-$350 k per year, and the effort began in 2002).

Section 2.4 Digital Preservation Testbed (Nationaal Archief)

9. The KRDS1 report also noted from the testbed project "...estimated costs c. 333 euros for the creation of a batch of 1000 records in the pre-archive phase. In contrast, once 10 years have passed and material has been transferred to an archive it may cost 10,000 euros to 'repair' a batch of 1000 records with badly created metadata "(Nationaal Archief, 2005a). These estimates of cost/benefits are probably an important observation on the impact of timing given outcomes from case study 1 in CMDP.

Section 3.1 OAIS Reference Model

10. Although agreeing with the comments on the OAIS Reference Model, it is perhaps an open question whether it is better to start from the OAIS Model or from other cost models. This is because, as noted in the report, OAIS itself is a relatively general and conceptual model and the degree of detailing of functions varies. This varying degree of detail may contribute to many of the observed deviations from OAIS in existing cost models. However, we do consider the OAIS reference model an appropriate starting point for CMDP, but only if specific implementations and enhancements to OAIS for costing purposes in ALM institutions are considered when necessary and duly documented. These may draw on other cost models where appropriate.

Other potential comparators

11. Although not as yet published, the Dutch Data Archiving and Networked Services (DANS) has been undertaking detailed cost modelling for its services over the past year using an economics undergraduate intern for a period of 9 months. The intern spent 3 months processing and preparing cost information (staff time sheets etc) and the remaining 6 months on data analysis. Results were presented at the Alliance for Permanent Access conference in November 2009 and the report is expected to be finalised early in the New Year.

Section 3.2 Method

12. This section follows a standard methodology for lifecycle activity based costing but focuses on specific activities (file format migration and preservation planning). Lifecycle costings model a lifecycle for a specific process(es) and then identify measurable

component activities, cost drivers (variables that affect the costs of the activity e.g. volumes, formats etc), and resources (staff time, equipment etc) to provide an understanding of costs for that process.

As noted by Gerlach in discussion of activity based costings for IT services (Gerlach 2002, p 64-5), a critical decision in a cost model's design is the defining of activities at an appropriate level of detail. This is because the choice of activity level greatly affects the accuracy and cost of developing and maintaining the model. Detailed activity modelling is usually needed for operations planning and process improvement, whereas more general high-level activity models are sufficient for cost management.

The CMDP model is currently developed at the detailed level for operations planning (future migrations) and process improvement. However the CMDP team may wish to consider the potential implications and utility of catering for dual application of a model in future. This is currently under consideration in KRDS2 with two "versions" of the KRDS2 model to be presented at different levels of detail for different purposes. An A4 overview of the KRDS2 consisting of just the main phases e.g. archive and sub-phases e.g. ingest has been produced for this purpose. A cost management application (sufficient to understand overall allocation of costs) can be obtained with a much lower overhead in terms of capturing the required data and may be helpful to some ALM institutions alongside options for more detailed methodologies and studies.

13. An additional application of activity based and lifecycle cost models is to help with the prediction of costs. This is a major function of CMDP and related models such as NASA CET and LIFE. To be successful this relies on documented costs from a set of identical or similar case studies to produce a costing for a new process (or a projection for a known documented process that is repeated with changed variables e.g. volumes). The accuracy for cost projections and their margin of error for a new process (e.g. future migration) is likely to be very dependent on the sample of comparators available. This may be a consideration for CMDP future projects – whether the number of comparators is sufficient and what is the acceptable margin of error for projections.

14. A major difficulty encountered by LIFE was the necessity of estimating and projecting for the number of file format migrations in future years given the absence of documented

data for this. This has a major impact on projected costs in its general preservation model. Similar challenges exist with this section of CMDP.

The issue of file format migration has been a major concern in digital preservation thinking ever since Jeff Rothenberg's 1995 Scientific American article "Ensuring the Longevity of Digital Documents" (Rothenberg 1995). More recently questions have been raised about the absence of mass file format obsolescence as predicted in 1995 by Rothenberg. David Rosenthal presented an important CNI Plenary presentation in April 2009 and subsequently in his blog (Rosenthal 2009). Although the reviewers would not agree with all of Rosenthals's conclusions, the key elements of his thesis are not really addressed or refuted by Morrisey (also on the blog) and we believe potentially have very important implications for preservation action costings in both LIFE And CMDP. In summary Rosenthal is arguing that for mainstream widely adopted file formats, software migration tools will be provided by the market and that the focus of digital preservation on file format obsolescence has been misplaced (although for a minority of file formats, preservation and obsolescence challenges will still be significant). He suggests that the real challenges are scale, cost (as a result of scaling), and intellectual property rights (legal costs of negotiation and for any challenges). If accepted this might suggest one future focus for CMDP might be on costs associated with scale such as workflow and bit preservation, and cost/benefit analysis for them.

15. Section 3.2.3 Cost factors notes the complexity of digital formats and structures (objects) has a significant influence on the cost of functional preservation by migration, and it is therefore important to be able to model this factor correctly. It also notes several attempts to define this complexity have failed, and Planets' conclusion seems to be widely accepted: The notion of "digital object complexity" has been disregarded as non-objective and non-scientific. The consultants cannot comment in detail on the CMDP alternate proposed approach to format complexity in this section as it is outside core competencies but we note DANS in their future work wish to undertake further work on the issue of "dataset complexity". Some related observations are also made in points 16-17 below.

16. The CMDP report allows for time spent reading documentation according to its length and complexity, but the quality parameter is still to be implemented in the model. Currently the model cannot make sufficient allowance for the worst case scenario (which occurs all too often), where there is no documentation at all especially after a number of years have passed. Lack of, or deficiencies in, documentation were a major cost in the UK Domesday project.

17. Additional comments from the perspective of a memory institution ICT department in the UK are provided in Appendix B. These partly cover the issue of curated databases in heritage institutions which have extended current life and will be many decades in "current use". Preservation in this context is very different from traditional archival workflows and preservation of non-current material. CMDP may wish to note this as a category of material that may be encountered in ALM institutions but is outside the CMDP model. It may also wish to consider the "Digital Continuity" approaches being developed by the UK National Archive to cater for "preservation" of current material over extended timeframes within government departments which is a similar issue. There are some potential applications of cost models to this type of activity.

Section 4 Resume of Danish Studies and Experience

18. This section is too brief. Consider adding a copy of the original questions or questionnaire and a table of responses. The number of potential case studies and cost datasets is a serious issue for viability of any follow-on and for the methodology.

Section 5 Case Studies

19. This was one of the most interesting and informative section of the report and illustrates the importance of being able to test and adjust models against documented actual experience. This is discussed in terms of implications for the methodology in points 12-14 above.

20. With respect to the final paragraph of p24/first paragraph of p25, it may be unrealistic to say that careful quality control is a pre-requisite for the cost model. This may limit its

application significantly. Another way of viewing this could be that there are additional cost drivers such as timing to be reflected in the model if it is to be applicable in such cases.

21. Discussion of Table 10. This is an important table and discussion. The points made in discussion on causes of variation in predicted results are all accurate but additional reflection by the project team on the results, such as what would be considered to be acceptable outcomes, and any specific changes to methodologies needed to achieve the acceptable outcomes (or even “can acceptable outcomes be achieved?”) would be valuable for planning future activity. We noted that outcomes improve if B & C migrations alone are considered but even within B and C the variation in predictions is in the range -28% to +62.5%. As part of the follow-up project to CMDP, it would be helpful to compare this directly with the range of variability experienced in NASA CET. This may allow you to see how comparable outcomes from the different models are and to quantify targets for any feasible improvements.
22. Section 5.5 Discussion of case studies page 26 – see comments 9, 12, and 20 above.

The Spreadsheet

22. This is detailed but very difficult to follow without more documentation, expansion of abbreviations used, and definition of terms. This will need to go beyond the existing documentation in the CMDP report and appendices and ideally the spreadsheet should be less dependent on separate reference to the documentation for use by new users. It would currently take more work before the aspiration that the user interface should be good enough that only a general knowledge of digital preservation, including information about the archive’s user groups and digital collections is needed to use the model.
23. Salaries may need definition as it can include basic salary, social costs, and pension contributions. There is a significant difference between basic salary and full salary costs to the institution.
24. The spreadsheet is currently bi-lingual. The majority is in English but I discovered several Danish terms (I think!).

5. Benchmarking against Related Approaches

While the CMDP project looked in detail at other cost approaches and how they could be integrated into the study, this evaluation process has carried out an additional independent benchmarking exercise against the related approaches in Keeping Research Data Safe (KRDS), LIFE, and the NASA Cost Estimation Tool (NASA CET) projects. The aim of the benchmarking has been to provide a brief overview of past and current work in the related approach, a comparison of it against the OAIS model (to give an independent and mutual point of reference), consideration of strengths and weaknesses from the perspective of CMDP, and finally some conclusions for each benchmark. Some benchmarking is also reflected in evaluator comments in the previous section above (section 4).

5.1. Keeping Research Data Safe

Introduction

[Disclosure: the evaluators are the lead organisation in this research project.]

The first Keeping Research Data Safe (KRDS1) study funded by JISC in the UK developed a cost model and indentified cost variables for preserving research data in UK universities. This was based on analysis of the OAIS, LIFE, and NASA CET models, additions by the project team, and a set of 4 case studies. The follow up to this study, Keeping Research Data Safe 2, commenced on 31 March 2009 and will complete in December 2009. The project has been identifying and analysing sources of long-lived data and develop longitudinal data on associated preservation costs and benefits. A costs data survey is expected to include 13-15 responses from UK and some international projects. There is also an extended “benefits framework” with two benefits case studies from Oxford University and the UK Data Archive; further review of the activity model in KRDS1; and discussion of costs data from Oxford, Southampton, the Archaeology Data Service, University of London Computer Centre, and the UK Data Archive. For further information see http://www.beagrie.com/jisc.php .

Strengths

- Has adopted a life-cycle costs approach and based its model on a rigorous examination of LIFE, NASA CET, and the OAIS model;
- Developed to integrate with full economic cost approach (“TRAC”) used by UK universities;
- Tested against real-life costs data and experience in a set of UK institutions;
- Introduces consideration of benefits alongside analysis of costs;
- Allows consideration of whole of lifecycle costs including pre-archive phase;
- Good at “whole of service provision” costing and cost management and broad brush trend data and cost factors.

Weaknesses

- Not implemented to detailed process level cost prediction spreadsheets and metrics to enable this;
- UK institutional context (but many generic features of wider relevance);
- Adapted for research data rather than ALM institutions (but again many generic features of wider relevance);
- Deviates from strict adherence to the OAIS model.

Comparison with OAIS

- KRDS has incorporated a pre-archive stage and pre-archive activities by the producer and archive (outreach) which we believe is a helpful addition;
- In common with LIFE, KRDS has a distinct stage for acquisition. Elements of this are only partially represented in the OAIS model. In addition it moves some functions from administration in the OAIS model and develops them as distinct activities in the Archive phase;
- KRDS adds a “first mover innovation” activity and disposal activity to the archive phase;
- There is otherwise broadly a close match between the Ingest, Archive Storage, Preservation Planning, and Access functions in KRDS and OAIS;

- KRDS adds a number of elements under Support Services and Estates appropriate to a cost model.

Conclusions

KRDS is less likely to be helpful to CMDP in its detailed process costing work. However the broader approaches and application to cost management and its consideration of benefits alongside costs may be of interest to CMDP. Similarly any cost data sets identified in KRDS2 could be of interest.

5.2. NASA CET

Introduction

The NASA CET is designed to provide NASA budget estimators, Principal Investigatorss, project managers, and resource planners with the capability to generate life-cycle cost estimates for implementing, operating and maintaining a science data system. The CET provides output in spreadsheet and graphical formats, and has various tools for what-if options, output review and manual override, parameter sensitivity tests, and creating new or updated historical project data in the Comparables Database (CDB). Users interface to the CET through Visual Basic developed forms that provide the inputs to Microsoft Excel where cost estimates and various output functions are generated. The Toolkit runs on PC or Mac platforms. The CET software package includes the CET, the Comparables Database (CDB), Users' Guide, and Technical Description Document. The current version is CET 2.1 and is available from http://opensource.gsfc.nasa.gov/projects/CET/CET.php .

Strengths

- The NASA CET has the most developed implementation and documentation of any of the cost models;
- It has adopted a lifecycle approach to costs and it can be mapped relatively easily into the LIFE and OAIS models;
- The model is derived from experience with 29 operational data centres from space and earth observation. This gives a strong empirical underpinning to the cost model and a strong degree of confidence in the statistical validity of its cost data for NASA activities;

- The NASA CET reference model has particularly good description of functions with definitions for Information Technology and systems costs associated with projects. Several of these relate to “Creation” phase activities excluded from the OAIS reference model;
- The CET has a set of 94 metadata fields (descriptors) with accompanying definitions which are used to describe specific functions. A number of these have menu options to capture key cost variables e.g. level of service or automation levels. There are sensitivity adjustments and linkages within the CET which inter-link components of the model and allow “what if” scenarios and ripple effects from changes in different elements to be modelled;
- The model distinguishes between “operational” and “support” functions. Support activities are essentially overheads that are distributed across one or more of the operating functions. Support activities may suggest the contours of a general institutional infrastructure that underpins a network of preservation activities;
- The reference model is supported by a prototype suite of Excel-based tools and a database of comparable costs from 29 projects/activities for estimating lifecycle costs;
- The cost estimation process currently has an overall average absolute error of 22.9%. As noted earlier (page 12, comment 21) it would be useful to compare the variability within this average to the CMDP. It could be argued that knowing how accurate your predictions are can be a strength (even if it is a broad margin), gained in this case from lots of longitudinal data. Comparative data and estimating techniques have been refined over time and it is believed the cost-estimation performance of the CET may now be as good as it can be.

Weaknesses

- The NASA CET model has been comparatively resource intensive to develop and implement. The level of investment has been roughly 2FTEs per year/$250-$350 k per year since 2002. This level of investment and resource is likely to be beyond ALM institutions;
- The CET model currently has no provision for long-term digital preservation costs or functions;
- It is based primarily on NASA projects and costs in space and earth observation research. As such it is particularly strong on major software development and support

that are needed in such projects. However this requirement will be at a much smaller scale in ALM institutions and less relevant to CMDP.

Comparison with OAIS

- NASA CET includes functions for Implementation, Sustaining Engineering, Technical Co-ordination, and Product Generation which are largely absent from OAIS;
- NSA CET has a more developed view of User Support than is present is OAIS;
- Archive and preservation activities are more developed in OAIS than NASA CET;
- There are relatively close matches between Ingest and Access/Distribution in both models;
- Engineering Support and Management/Administration match but have significant variations in level of detail and coverage;
- Data Management is treated as a separate function in OAIS but not in NASA CET.

Conclusions

The NASA CET is the earliest and most evolved cost model under consideration. Although the least relevant in some ways to ALM institutions, it is the only model to take a rigorous statistical approach to cost prediction and to define and refine its levels of accuracy. Its approaches to this and implications from it are highly relevant to CMDP. The implementation (CET spreadsheet, comparators database, and user documentation) may also provide pointers for future development of CMDP although levels of available resource will not be comparable and relevant adjustments would be necessary.

5.3. LIFE

Introduction

LIFE (Life Cycle Information for E-Literature) is a collaboration between University College London (UCL) and the British Library. The Project has developed a methodology to model the digital lifecycle and calculate the costs of preserving digital information. LIFE completed its first phase in April 2006 and the second phase in August 2008.The third phase of the project commenced in August 2009 and will complete in July 2010. The aims statement for LIFE 3 is as follows:

"By producing a predictive costing tool, LIFE3 will significantly improve the ability of organisations to plan and manage the preservation of digital content. The project will expand its existing Generic Preservation Model to create a comprehensive suite of models covering all life cycle stages, providing greater accuracy and assurance in estimation. The predictive costing tool will be made available towards the end of 2010, as both a web application and an Excel-based model."

It is too early to assess LIFE 3 and comments below relate mainly to LIFE 1 and 2.

Strengths

- LIFE focuses primarily on the library sector and library materials;
- The LIFE projects have adopted a lifecycle approach to digital preservation costs. This has a long pedigree in cost modelling in other sectors; has been applied to costing in traditional library collection management; and has been advocated as an approach to digital collection management and costing digital preservation;
- LIFE has added an optional stage for Creation or Purchase on to the OAIS model. This is a helpful development and could allow a model to reflect dependencies and implications for costs between pre-archive and post-archive phases;
- LIFE draws heavily on the experience of the British Library which has preservation as a core function. This experience is supplemented by experience from other institutions;
- LIFE allows for more pro-active collection development processes than the OAIS model. In particular it has an Acquisition stage with a selection element which will be more appropriate for some organisations with a degree of choice over acceptance of data collections offered for ingest;
- LIFE 3 aims to develop detailed process level calculations and extend its Generic Preservation Model. This may be of particular relevance to CMDP.

Weaknesses of the LIFE Model

- LIFE has arguably less user input and end-user focus than other cost model projects;
- Many organisations need to consider Full Economic Costs (FEC) and the LIFE 2 model focuses mainly on lifecycle episodes rather than associated ongoing and support infrastructure costs. Treatment of some direct and most indirect costs is poorly developed in LIFE 2 but may be addressed in LIFE 3;

- The LIFE 2 model tends towards an implicit assumption of a uniform preservation aim or outcome and therefore cost;
- There has been no real testing of the estimation process in the generic cost model against actual cost outcomes and the model generates figures that have a potentially misleading appearances of accuracy with no data on the likely accuracy of the cost predictions;
- LIFE 2 currently treats costs in each stage as independent elements. In practice, in many cases choices made at one point in the lifecycle could ripple across to other stages and costs. Greater sensitivity may need to be built in to the model to these choices, linkage between different elements, and modelling how costs are affected.

Comparison with OAIS

- LIFE has incorporated an optional pre-archive stage for creation or purchase which we believe is a helpful addition;
- LIFE has a distinct stage for acquisition which we would also support. Elements of this are only partially represented in the OAIS model;
- LIFE has a separate stage for Metadata – we prefer the OAIS emphasis on both documentation and metadata (descriptive information) and leaving metadata in situ within the lifecycle;
- There is broadly a close match between the Ingest, Archive Storage, Preservation Planning, and Access functions in LIFE and OAIS;
- The Data Management, Administration and Common Services functions are more developed in OAIS than LIFE;
- LIFE adds a number of elements under Economic Adjustments appropriate to a cost model.

Conclusions

The LIFE project is potentially the closest to CMDP in terms of intended audience and methodology. However LIFE 1 and LIFE 2 have a number of weaknesses which severely reduced their value for CMDP. Much will depend of how LIFE 3 unfolds and its next stage should be closely monitored by CMDP.

6. FUTURE OPPORTUNITIES

1. **Student study placements.** A considerable challenge for any cost model is the requirement for good data on costs and a reasonable sample of comparable situations on which predictive models can be based and which can help refine cost factors and their relative importance in different situations.

   Even where costs data such as staff timesheets exist, considerable time is needed to process and prepare that data for any analysis to fit the model.

   We have noted that DANS (see section 4.2 comment 11 above) successfully used an economics undergraduate intern for a period of 9 months in its costs study. We suggest CMDP might also consider a number of student placements in any future stage as a means of preparing and/or creating and analysing cost data sets in a number of Danish ALM institutions. This might be a particularly useful way of re-inforcing further engagement with the CMDP target user community and be a relatively cost-effective mechanism for them.

2. **Refining the accuracy of the prediction model.** As noted above we believe some reflection will be needed on the acceptable and/or achievable level of accuracy of the model. Generating further cost data sets as a first step will be helpful in providing additional input into refining the accuracy of the model.

   It is also possible that LIFE 3 will produce tools and further refinements of its methods for the General Preservation Model which could be used or inform the CMDP. LIFE 3 should produce its outputs in late 2010.

3. **Refining Format Complexity**. We have noted CMDP's interest in further defining measures for format complexity and their impact on costs. It is possible that LIFE 3 will include work in this area useful to the CMDP and we have also noted that DANS proposes further work on dataset complexity for its cost model. Both of these projects may produce useful outputs or sources of collaboration for CMDP.

   Other institutions will have extensive experience of format migrations and a close interest in costs for them might also be very helpful. Possibilities might include

Portico or the Internet Archive internationally and contacts with Denmark with producers and contractors involved in migration and normalisation of records within Government.

CMDP has already presented its preliminary work as a conference paper at iPRES 2009. We suggest the topic of Format Complexity and Digital Preservation costs might be a suitable topic for a follow-up and more detailed discussion in a workshop at iPRES 2010 in Vienna or at a SUN PASIG meeting (often held in parallel to iPRES). Other relevant organisations such as LIFE, DANS and Portico might be participants with CMDP. A "white paper" may help to focus discussion and review comments.

4. **Extending work to include Cost/benefit Analysis.** We suggest the next phase of CMDP inter alia might focus on quantifying benefits alongside costs for a specific case study. This will be an important message to develop for funding agencies as well as participants. CMDP may find elements of the work in the forthcoming KRDS 2 project report helpful for this.

5. **Developing dual applications of the Model.** We have noted (section 4.2 comment 12) how Gerlach and KRDS2 suggest activity based cost models can be applied at two main levels of detail and for different purposes. The CMDP team may wish to consider the potential implications and utility of catering for dual application of a model in future. This is currently under consideration in KRDS2 with two "versions" of the KRDS2 model to be presented at different levels of detail for different purposes. A cost management application (sufficient to understand overall allocation of costs) can be obtained with a much lower overhead in terms of capturing the required data and may be helpful to some ALM institutions alongside options for more detailed methodologies and studies. We suggest CMDP might also consider it and KRDS 2 could provide helpful information and models for this.

6. **Exploring Scale issues.** Rosenthal has proposed that scale is one of the major issues for digital preservation. If accepted, this might suggest that one future focus for CMDP might be on costs associated with scale such as workflow and bit preservation, and perhaps linking this to cost/benefit analysis for them.

7. **“Digital Continuity”.** Appendix B covers the issue of curated databases in heritage institutions which have extended current life and will be many decades in “current use”. Preservation in this context is very different from traditional archival workflows and preservation of non-current material. The “Digital Continuity” approaches being developed by the UK National Archive to cater for “preservation” of current material over extended timeframes within government departments is a similar issue. There are some potential applications of cost models to this type of activity which might be a future application for CMDP.

8. **Collaboration with other Projects.** We note above the potential for collaboration with other projects. This might be extended to funded partnerships where funding conditions allow, as well as mutual in-kind exchange of experience. Joint funding bids might also be a possibility to tap into non-national funding such as the European Commission programmes or independent foundation funding.

7. Conclusions

The CMDP has produced interesting results with relatively modest resources and should be continued to build on this initial work. We believe there are a number of potential areas that could profitably be explored in any follow-up phase.

In addition, there are also a number of approaches that could leverage its work, including use of student placements and potential collaborations with related projects that we have identified elsewhere in the evaluation report.

8. References

Consultative Committee for Space Data Systems (2002) *Reference Model for an Open Archival Information System (OAIS)*, CCSDS 650.0-B-1, Blue Book, 2002, (ISO14721:2003). http://public.ccsds.org/publications/archive/650x0b1.pdf

Gerlach, J., Neumann, B., Moldauer, E., Argo, M., and Frisby, D. 2002. Determining the cost of IT services. *Commun. ACM* 45, 9 (Sep. 2002), 61-67. DOI= http://doi.acm.org/10.1145/567498.567500

Kejser, U., Nielsen, A., & Thirifays, A. (2009a) *The Cost of Digital Preservation Project Report v.1.0*, November 2009

Kejser, U., Nielsen, A., & Thirifays, A. (2009b) *Cost model for digital preservation – all elements, using OAIS functional model.xls*

Kejser, U., Nielsen, A., & Thirifays, A . (2009c) *Cost Model for Digital Curation: Cost of Digital Migration*. Paper presented at The Sixth International Conference on Preservation of Digital Objects.

Kejser, U., Nielsen, A., & Thirifays, A. (2009d) *Cost Model for Digital Curation: Cost of Digital Migration*. iPRES 2009 PowerPoint presentation. Retrieved 4 December from: http://www.cdlib.org/iPres/presentations/Kejser.pdf

Kejser, U., Nielsen, A., & Thirifays, A. (2009e) *Appendix2_OAIS Analysis-2*

Nationaal Archief, 2005a, *Costs of Digital Preservation version 1.0 May 2005* (Digital Preservation Testbed, The Hague, Netherlands). Retrieved from http://www.digitaleduurzaamheid.nl/bibliotheek/docs/CoDPv1.pdf

Rosenthal, D., 2009, *Spring CNI Plenary: The Video* http://blog.dshr.org/2009/07/spring-cni-plenary-video.html and blog entries *Spring CNI Plenary Re-Mix*

http://blog.dshr.org/2009/04/spring-cni-plenary-remix.html and *Sheila Morrissey's Comment* http://blog.dshr.org/2009/05/sheila-morrisseys-comment.html

Appendix A: Self-Evaluation Analysis

	Principal Goal of CMDP Project
1.	**The project has as its goal to set up a generic model for the calculation of costs related to preservation of digital materials at ALM institutions.** A model has been created which addresses migration and preservation costs. It has been tested on two case studies. The final outcome however was limited in that the model only covers functional preservation by migration (Kejser et. al (1), 2009, p 8, paragraph 3).
	Success Criteria of CMDP Project
2.	**Investigation of the LIFE Project and its follow-up [LIFE2] and related cost models.** A thorough investigation and analysis of the LIFE project [1 and 2] was made (Kejser et. al, 2009a, Section 2.1). The Life 'Costing model' and 'Generic Preservation Model' were analysed in detail, as well as usefulness for the project. Investigation of related cost models included KRDS; NASA CET and the Dutch Digital Preservation Testbed (Kejser et. al, 2009a, Section 2.2 – 2.4). 'Usefulness for the project' from these were analysed among other factors.
3.	**Development of a cost model which is tested in detail on Danish data and experience**. Completed but not for all the functional entities in OAIS. Two case studies have been carried out on which the model developed so far has been tested (Kejser et. al, 2009a, p 4, paragraph 2). However, the model is not yet tested on completely independent data (Kejser et. al, 2009a, p.3, paragraph 7) such as collections within other ALM institutions (Kejser et. al, 2009a, p 7, paragraph 2).

4.	**Use of the cost model on the estimated amounts of digital material which must be preserved for the state archives, libraries and museums in the next ten years.** This remains unfinished, primarily because within the time limits it has not been possible to obtain and standardise the desired information from the cultural heritage institutions. (Kejser et. al, 2009a, p 4, paragraph 2)
	SUB-GOALS OF CMDP PROJECT
5.	**The cost model should be able to deal with all the types of materials that the ALM institutions must preserve and calculate costs for relevant preservation strategies such as migration and emulation.** The model is a good foundation for further development and operationalising of the remaining function areas. The final outcome however was limited in that the model only covers functional preservation by migration (Kejser et. al, 2009a, p 8, paragraph 3). Model not yet tested on independent data (Kejser et. al, 2009a, p.3, paragraph 7) or on collections within ALM institutions (Kejser et. al, 2009a, p 7, paragraph 2)
6.	**The cost model must be based on a well defined breakdown of all the activities that are encompassed by digital preservation.** Functional breakdown of OAIS carried out and identification of cost critical activities. Mutual dependencies analysed and expressed in formulas (Kejser el al 2009e). By including the three OAIS roles of Producer, Consumer and Management, the project has ensured that the model also takes into account the external cost factors that affect the OAIS archive. (Kejser et. al, 2009a, p 28, paragraph 2)
7.	**The model must cover both the costs of the archive systems themselves and the costs related to the flow of data through the OAIS archive**. For this purpose the project defined a cost critical flow between the relevant functions within the OAIS archive. (Kejser et. al, 2009c, p 2, paragraph 6). The flows in and out of each function have been assessed as to which are cost critical. Two activities are identified as cost critical: Monitoring of Technology (emerging digital technologies, information standards and computing platforms) and Generation of reports (reports, external data standards and

	technology alerts). (Kejser et. al, 2009a, p 16, paragraph 4). A number of OAIS components are not relevant for the CMDC and have thus been excluded; others have been combined. (Kejser et. al, 2009c, p 5, paragraph 3)
8.	**The model must be able to take into account outside factors which influence the institution's costs, e.g. overall policies and budgets; technological developments and changes in the behaviour of different user groups.** The model has to handle dependencies better, because no cost critical activities stand alone. Their mutual implications are difficult to account for, but highly cost sensitive. The most obvious example from case 1 was the model's difficulty of calculating the high cost of the Migration Plan phase (Kejser et. al, 2009a, p 27, paragraph 1). When used for estimating future cost the precision is even more uncertain due to the challenges posed by handling the predictive element.
9.	**The institutions must be able to use the results from the cost model for strategic decision making and budgeting.** Model underestimated the large project management costs for one migration batch. For another set of data it overestimated the costs of the system and design of new data (Kejser et. al, 2009a, p 3, paragraph 8)
10.	**The model should be based on well defined accounting principles and include all costs, including both investment costs and operating expenses, direct as well as indirect** CMDP is not yet operational in relation to managing investment costs or direct and indirect running costs (Kejser et. al, 2009a, p 30, paragraph 3). Overhead covers indirect costs, i.e. indirect staff, facilities, general management and administration (Kejser et. al, 2009c, p 2, paragraph 5)
11.	**The model must be able to measure actual baseline costs, e.g. cost based on experiences (ex-post) but the activity based approach also allows tracking cost over time and estimating future cost (ex-ante), ideally within a 10 to 20 year perspective** The CMDC is applicable for measuring actual baseline costs, i.e. cost based on experiences (ex-post), but the activity based approach also allows tracking costs over time. The precision of the model is low, but regarding the exact degree we dare not make any conclusions. When used for estimating future cost (ex-ante) the precision is

	even more uncertain due to the challenges posed by handling the predictive element, which influence various aspects of the model. (Kejser et. al, 2009c, p 5, paragraph 4)
12.	**The model must therefore also be able to manage financial adjustments, such as inflation and deflation, interest and discounting.** It was roughly assumed that on average preservation formats will be usable for 10 years, and every 5 years a migration is performed, migrating half of the content. Considering the many different types of formats and the ability to calculate the cost of a migration that most of all resembles a normalisation, the model also needs more parameters to reflect that not all preconditions are fulfilled (Kejser et. al, 2009a, p 27, paragraph 1)
13.	**The cost model should be able to be used to compare costs between the preserving institutions and this requires that the quality of the preservation is also comparable.** The conclusion of the above test with cost data is that CMDP manages very well, if the prerequisites for the use of the model are fulfilled. But this also means that there is a significant responsibility to inform users of the model's prerequisites.(Kejser et. al, 2009a, p 27, paragraph 3)
14.	**The user interface should be good enough that only a general knowledge of digital preservation, including information about the archive's user groups and digital collections is needed to use the model** n/a
15.	**Examination of Danish studies and experience and adjustments of the model to these** The project concluded that there is no Danish cost model and only sparse systematically documented expenses for projects within digital preservation. (Kejser et. al, 2009a, p 22, paragraph 2)
16.	**Collection of information on holdings and the expected growth in digital materials.** Work in progress
17.	**Application of the model to Danish conditions (based on the collected information).** Case studies carried out, but model not yet tested on independent data (Kejser et. al, 2009a, p.3, paragraph 7) or on collections within ALM institutions (Kejser et. al, 2009a,

	p 7, paragraph 2)
18.	**A general cost methodology for digital preservation, including a function model based on the OAIS Reference model and a set of generally accepted accounting principles.** By including the three OAIS roles of Producer, Consumer and Management, the project has ensured that the model also takes into account the external cost factors that affect the OAIS archive. (Kejser et. al, 2009a, p 28, paragraph 2) The basic formula for an activity is the effective time required to complete an activity (measured in pw) times wage level (including overhead) plus purchases (monetary value). (Kejser et. al, 2009c, p 2, paragraph 5)
19.	**A cost model for functional preservation by the preservation strategy digital migration** . Note that the flow does not yet include the cost critical activities of requesting the content to be migrated from Archival Storage, nor ingesting the new Information Package (IP) version back into Archival Storage. (Kejser et. al, 2009a, p 16, paragraph 3) the CMDC does not presently reflect the size of migration projects well enough (Kejser et. al, 2009c, p 4, paragraph 5). In order to calculate the time it takes to execute the actual migration, we have introduced a Processing factor. It depends on the Interpretation factor, the amount of data, the computer power, and on the number of computers (Kejser et. al, 2009c, p 3, paragraph 5) The way the migration project classified the tasks does not correspond to the way it has been done in CMDP (and therefore not in OAIS either), and thus it was necessary to map the cost data between the migration project and the CMDP. (Kejser et. al, 2009a, p 23, paragraph 8) Another interesting fact is that the cost data shows that it is equally expensive to make migration plans and develop software, while the CMDP underestimates the cost of the Migration Plans step. (Kejser et. al, 2009a, p 25, paragraph 3)
20.	**A set of formulas which can transform input to cost data.** In an attempt to tie the factor to more measurable components we have for example introduced the Format Interpretation factor, which relates the complexity to certain characteristics of the

	formats' documentations. (Kejser et. al, 2009a, p 17, paragraph 1) The CMDP is designed to allocate a certain number of 'pw' to the IP designing process. This teaches us that if data comply with the IP design at hand, the model should exclude this cost. (Kejser et. al, 2009a, p 25, paragraph 2)
21.	**A tool set (including documentation of the calculations), which makes it possible for the user to carry out automated cost calculations in relation to functional preservation**. To increase precision of the estimates, the migration tool development was divided into three modules, namely a Reader for the source format, a Writer for the destination format and a Translator to map between the two formats, each tool based on a calculated formula (Kejser et. al, 2009a, p 19, paragraph 2).The tools were then applied to the case studies to automate the process, where possible: for Case 1, automation only reached 20% (Kejser et. al, 2009a, p 25, paragraph 2), whereas with Case 2, automation was at 99%. However in some cases, automation of migration is dependent on the quality or 'compliance' of the data (Kejser et. al, 2009a, p 26, paragraph 2).

APPENDIX B: EXAMPLE OF UK HERITAGE BODY ICT PRACTICE

The work of the ICT department is primarily concerned with maintaining the mechanisms for access to current datasets rather than preserving complete datasets which are unchanging. "Old" data is generally preserved by migration and merging into these current datasets, and they have digital material dating back to the mid-1980s which forms part of these datasets. Therefore, their experience may not correspond completely with the projects described in the report. However they are likely to correspond with some Danish ALM institutions whose principal digital assets are "curated datasets" which are constantly updated but incorporate "preserved" material and might therefore provide an additional perspective.

We do not have a specific monitoring tool or function as described in the report. However, in the normal course of our work we need to be aware of new versions of software etc, especially as older versions move out of support from the suppliers, and for us this is one of the drivers for migration.

The main drivers for migration generally occur before the point of format or media obsolescence.

1. Software version updates (applying both to database versions and operating systems):
    a. New version providing improved functionality
    b. Older versions moving out of support
    c. Organisational policies towards standardisation of software.
2. Hardware changes
    a. Servers are generally leased, so it makes sense to request upgrades to newer servers with better performance and capacity at similar cost.
3. Business changes in the organisation which may dictate merging of datasets e.g. to improve public access and increase usage.

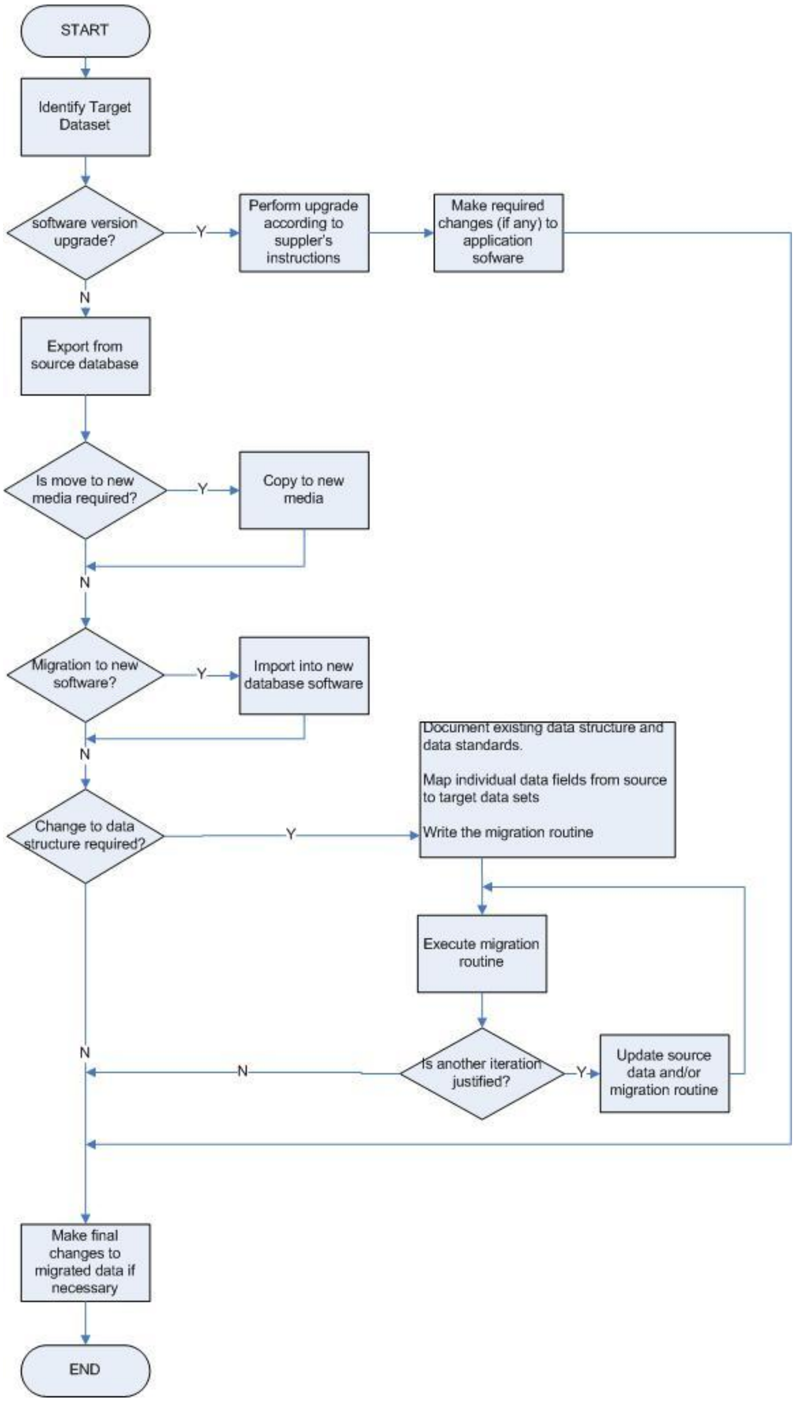
The process of migration is detailed in the flowchart below.

We find the split between reader, writer and translator somewhat artificial, unless this is just a question of semantics; they might relate to export, import and migration on our flowchart.

Simple upgrade to the next version of a commercial software package is usually not too difficult or time-consuming, but if this is a major upgrade or intermediate upgrades have been missed out then this increases the complexity. For us, the hard work can be more in the adaptation of the applications and access tools (search mechanisms, forms, reports, websites etc) than in the data itself.

The tables in the report don’t include database formats; this assumes you have already used an export facility to produce csv (txt) or xml.

Re. the number of machines used. We don’t know how useful this concept is for us, as you can only use 1 server at a time for a single dataset.


START
Identify Target Dataset
software version upgrade?
Y
Perform upgrade according to suppler's instructions
Make required changes (if any) to application sofware
N
Export from source database
Is move to new media required?
Y
Copy to new media
N
Migration to new software?
Y
Import into new database software
N
Change to data structure required?
Y
Document existing data structure and data standards.
Map individual data fields from source to target data sets
Write the migration routine
Execute migration routine
Is another iteration justified?
Y
Update source data and/or migration routine
N
N
Make final changes to migrated data if necessary
END