Project no. 269977

APARSEN
Alliance for Permanent Access to the Records of Science Network

Instrument: Network of Excellence

Thematic Priority: ICT 6-4.1 – Digital Libraries and Digital Preservation

D32.2 Report on Testing of Cost Models and Further Analysis of Cost Parameters

Document identifier:	**APARSEN**-REP-D32_2-01-1_0
Due Date:	2013-06-30
Submission Date:	**2013-06-30**
Work package:	WP32
Partners:	BL, CERN, DNB, KNAW-DANS, STFC
WP Lead Partner:	BL
Document status	Released
URN	urn:nbn:de:101-20140516232

Abstract: This report presents the results of the analysis of cost parameters and the testing of cost models used within digital preservation (DP) both for services and repositories. The relationship between costs and benefits is reviewed in the context of DP and the links to an EU Co-ordination Action are also considered.

Delivery Type	Report
Author(s)	Kirnn Kaur (KK) - BL, Robert Darby (RD) – STFC, Patricia Herterich (PH) – CERN, Karlheinz Schmitt (KS) – DNB, Sabine Schrimpf (SS) – DNB, Heiko Tjalsma (HT) - KNAW-DANS, Simon Lambert (SL) – STFC
Approval	Simon Lambert, David Giaretta
Summary	Analysis of cost parameters and testing of cost models for digital preservation
Keyword List	Cost models, cost parameters, ISO16363 Audit and certification of trustworthy digital repositories, digital preservation
Availability	☒ PUBLIC

Document Status Sheet

Issue	Date	Comment	Author
0.1	2013-05-30	First draft	KK, RD, PH, HT, KS
0.2	2013-06-24	Final draft for review	KK, KS, PH, HT
0.3	2013-06-26	Peer review (external)	Neil Grindley (JISC)
0.4	2013-06-26	Peer review (internal)	Ashley Hunter (Tessella)
0.5	2013-06-27	Final updates	KK, RD, PH, HT
1.0	2013-06-28	Final report	KK

Project information

Project acronym:	**APARSEN**
Project full title:	**Alliance for Permanent Access to the Records of Science Network**
Proposal/Contract no.:	**269977**

Project coordinator: Simon Lambert/David Giaretta

Address:	STFC, Rutherford Appleton Laboratory Chilton, Didcot, Oxon OX11 0QX, UK
Phone:	+44 1235 446235
Fax:	+44 1235 446362
Mobile:	+44 (0) 7770326304
E-mail:	simon.lambert@stfc.ac.uk / david.giaretta@stfc.ac.uk

EXECUTIVE SUMMARY

This report presents the results of the analysis of cost parameters, the testing of cost models, the relationship between costs and benefits and the links to an EU Co-ordination Action. An in-depth analysis of the mapping of cost parameters to the ISO 16363 Standard on audit and certification of trustworthy digital repositories identifies and focusses on areas for further investigation and development. The results are presented following a gap analysis to provide indications of where future assessments and reviews could be undertaken. This is particularly valuable where cost models are still under development or where further projects are undertaken, for example, the Coordination Action, 4C (Collaboration to Clarify the Costs of Curation). Three cost models were tested providing an insight into how effective cost data, from digital repositories or archives, is in providing the costs of digital preservation services or workflows when applied to different cost models. Recommendations are given on any future developments of the three cost models with advice on the creation of new models. A review of costs in relation to benefits is presented. Finally, the links to the 4C project are presented where the outputs of this research will be further developed.

CONTENT

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 APPROACH AND METHODS

1.1.1 Analysis of cost parameters

The definition of a cost parameter could be regarded as: a unit of measure to define a cost associated with a process or activity; can be fixed or variable, direct or indirect.

As stated in the previous deliverable, D32.1 [1], the International Standard on Audit and Certification of Trustworthy Digital Repositories (ISO 16363) [2], which is based on the Open Archival Information System (OAIS) reference model (ISO 14721) [3], was selected to be used as a common denominator between all cost models. The Standard (ISO16363) provides a benchmark which allows at least a certain level of comparability between the various cost models and also provides a framework for mapping the cost parameters by cost model to allow for comparisons to be made.

To re-cap, the International Standard on audit and certification of trustworthy digital repositories is divided into five sections. The first two sections provide the overall document structure and an overview of audit and certification criteria. These two sections are excluded from the mapping exercise. Sections 3 to 5 provide metrics or measures which allow for the audit and certification of a trusted digital repository, namely; Organisational Infrastructure, Digital Object Management and Infrastructure and Security Risk Management. These sections as well as associated sub-sections are shown in Figure 1 below:

<u>ISO16363: Audit and certification of trustworthy digital repositories - extract</u>

3. Organisational infrastructure:
   3.1. Governance and organisational viability
   3.2. Organisational structure and staffing
   3.3. Procedural accountability and preservation policy framework
   3.4. Financial sustainability
   3.5. Contracts, licenses and liabilities

4. Digital Object Management
   4.1. Ingest: Acquisition of content
   4.2. Ingest: Creation of AIP
   4.3. Preservation planning
   4.4. AIP preservation
   4.5. Information management
   4.6. Access management

5. Infrastructure and security risk management
   5.1. Technical infrastructure risk management
   5.2. Security risk management

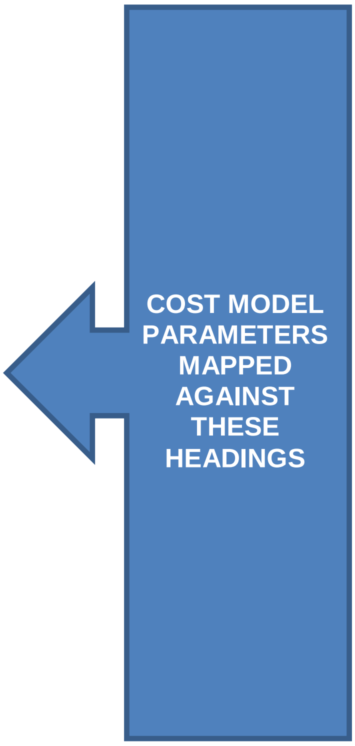


Figure 1: Extract of ISO16363

There are two levels of mapping: at sub-section level e.g. 3.1, 3.2 or at the level below this e.g. 3.1.1, 3.1.2. We initially focussed on the mappings at the sub-section level, see Figure 2 below for an extract of the standard. For each section in the detailed level below, the cost parameters were reviewed further where relevant comparisons could be made. In contrast to D32.1, the gap analysis presented in this deliverable is based on sub-section level as well as on this more detailed level (called further level in Figure 2). Not all parameters were sufficiently defined to allow for mapping at this detailed level. In some cases definitions were missing, and as stated later in this report, cost parameter definitions were identified as an important area for further investigation.

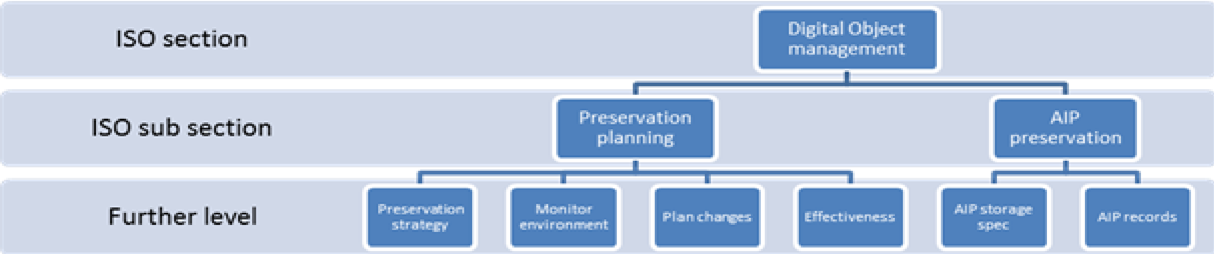


Figure 2: Sections from ISO16363

A requirement of the exercise was that all parameters should be mapped. Where parameters could be categorised across a number of sections, details were provided in the comments column as well as being re-entered against the relevant heading. Where mappings were implied or details were ambiguous, this information was also provided. ANNEX 3 of the previous report, D32.1, provides a table of the detailed sections extracted from the standard for the mapping exercise.

1.1.2 Testing of cost models

Cost models were selected, from those already analysed, to be tested with data collected from within the consortium. For the selection criteria, it was important that there was in-depth knowledge and understanding of the cost model. The ideal situation would be if a cost model owner or creator was involved and for this reason the DANS and DP4lib cost models were selected. The LIFE3 model was the final model selected for testing as the data and case studies used to set up the model were available to one of the partners involved in this work.

The task of collecting cost data was undertaken. The aim was to test models against real data, where possible, for cost data specifically related to the costs of digital preservation. The first step was to assess whether APARSEN partners could provide some examples of estimated or actual costs of preservation which might arise from:

- project proposals to a funding body where an estimate had been made for digitisation or preservation costs (such an estimate might itself have been based on a model, of course)
- experimental data and estimates of costs of digital preservation activities e.g. storage or archiving websites
- costs associated with particular preservation actions such as format migration, or particular changes in the environment such as a change in the designated community
- budgets or actual costs incurred as produced for management reports
- historical experience with running costs of repositories or archives

Data confidentiality issues and concerns were to be addressed by ensuring the anonymisation of the data collected. The cost data (estimated, allocated or actual) related to digital preservation or digitisation processes or activities was to be manipulated through allocation and apportionment methods to ‘best-fit’ the cost model being tested.

Initially, the data collection exercise did not provide the relevant data required to enter across the three models selected for testing, which would have provided comparative analysis of the outputs produced by each model. Data was not available for a number of reasons; organisations were unable to extract relevant information related specifically to digital preservation (DP) activities. The idea of trying to extract costs purely related to DP activities proved to be very difficult as most of the costs simply were not available. Organisational structures and the coding of financial data within organisations proved to be a barrier in extracting the data. We also tried to obtain financial data related to digitisation projects. However, there were commercial confidentiality issues related to the costs, for example, internal charging for digitisation or third party services.

As a contingency we looked at the cost data which had been used to set up, test and validate the models selected for testing. This data was extracted and used to carry out testing. The aim of the exercise was to test the workability of a model or its flexibility for adaptation to other services and workflows. By testing how data not previously used in setting up the cost model would be transferable to another model this would provide insights into how 'usable' the model would be to those wishing to cost DP activities or workflows in their own organisations. In order to provide comparisons of how the data matched across the various models, as much of the test data as possible from each provider was to be entered in to the alternative models, using relevant apportionment and allocation methods to distribute the data values as required. By analysing the areas where data entry was not possible, valuable insights would be sought to inform the development of cost models by their owners as well as for those wishing to set up their own cost models.

1.1.3 Modelling the benefits of Digital Preservation in relation to costs

Various ways were considered in which the relationship between the costs of preservation and its benefits have been described and modelled in published literature on cost models. Some of the main challenges of modelling benefits and proposed approaches in the literature on data preservation are presented. This is followed by analytic summaries of three key attempts to conceptualise and model the digital preservation cost-benefits relationship: from the Blue Ribbon Task Force, the Keeping Research Data Safe projects, and the DANS Balanced Scorecard method. In conclusion we consider how the benefits of digital preservation might be articulated in relation to the cost model within the overall context of the data preservation business case, and make suggestions for further development of business models and case studies in digital preservation.

2 ANALYSIS OF COST PARAMETERS

2.1 INTRODUCTION

The preliminary findings for this exercise were first reported on in February 2013 and presented in D32.1 [1]. The cost mapping exercise has now been finalised and was carried out successfully for eight of the ten cost models listed in Table 1 below. Further details of these cost models are provided in ANNEX 1 of this report.

Table 1: Cost models included within cost parameter mappings exercise

Cost Model	Mapping status	Level of mapping analysis
CET	Found to be too technical	Not mapped
CMDP	Completed	Detailed headings
DANS	Completed	Detailed headings
DP4lib	Completed	Detailed headings
ENSURE	Completed	Detailed headings
ISIS	Completed	Detailed headings
KRDS	Completed	Detailed headings
LIFE3	Completed	Detailed headings
OECD	Found to be not relevant	Sub heading
PrestoPRIME	Completed	Detailed headings

This chapter presents the findings of the mapping exercise with a focus on the analysis of areas for further investigation and development.

2.2 FINDINGS OF COST PARAMETER MAPPING EXERCISE

2.2.1 Cost parameters mapped

For three cost models: DANS, DP4lib and LIFE3, the parameter mappings were already successfully completed at sub-section level in deliverable D32.1. The results presented here include further analysis at the detailed level. CMDP, KRDS and PrestoPRIME models, only partially mapped in the previous deliverable, D32.1, and are now finalised. Two further cost models: ENSURE and ISIS were included in this phase of analysis. The models are presented in alphabetical order.

- CMDP – It is important to note that this cost model is not fully developed. The cost model is divided into the following sections: ingest, preservation planning, administration, data management, access, common services, and archival storage. The only areas currently populated with cost parameters are ingest, archival storage, administration and preservation planning. The other areas still to be developed are data management, access and common services. Where parameters are available, all high level parameters could be mapped. Some, however, were mapped based on certain assumptions due to a lack of detailed cost parameter definitions or they were mapped based on certain sub-parameters and therefore the match can be considered as rather vague. For the parameters where no definitions were available, these were mapped based on assumptions related to their title and any other relevant information available. When this cost model is developed further, it may be worth reviewing the cost parameter mappings based on the updated information. This is outside the scope of this project. Overall, the CMDP model is recommended for costs in the fields of Digital Object Management and Infrastructure and Security Risk Management.

- DANS – Mapping the cost parameters across the sections of the ISO standard was very difficult. In particular most headings in the section Organisational Infrastructure do not have any cost parameters mapped to them. A substantial number of these headings are more or less covered but they are not in any way quantified in the DANS cost model. Approximately half of the standard headings were not covered and do not seem to fit into the structure of the cost model. These areas may not be relevant to this cost model and will not be reviewed further. Many of the headings not covered in the model relate to Risk Management actions and obligations like the recording of error reports regarding authenticity for example (ISO ref 4.6.2). Risk Management is very much in development at DANS at the moment. It is the intention to develop the DANS cost model further, due to the dynamic nature of the data archive which is changing continuously over time, in which case the ISO headings not covered now might be covered in the future due to the work of this activity.
- DP4lib – Cost parameters could not easily be mapped to the ISO standard. However, a lot of sections could be covered, except for preservation actions regarding technical risk management and security risk management. However, a detailed specification of the underlying services and actions will be undertaken and cost parameters are currently in development for Security Risk Management. Furthermore, governance and organisational viability are out of scope of the cost model as for the cost model owner, the German National Library, this area would be covered due to its statutory obligations anyway. The DP4lib model is based on the activities implemented in the institutional workflow of the cost model owner and therefore still has some gaps as no migration or emulation activities covering some essential preservation actions are implemented within DP4lib services. On the other hand, the model is still under development and areas for further development have been identified, for example, measurement methods covering preservation planning activities. Overall, the DP4lib cost model can be recommended for covering most of the areas of the standard.
- ENSURE – Most of the low level cost parameters were mapped based on the assumption that where parameters partly match the standard they were mapped to the nearest relevant heading. Therefore, mapped parameters do not necessarily cover all areas within the heading. Some parameters could not be mapped as they did not seem to be defined (delivery response cost, information package retrieval cost). Overall, the ENSURE model can be recommended to cost Digital Object Management activities.
- ISIS – Mappings are based on the assumption that where parameters partly match the standard they were mapped to the nearest relevant heading. Some parameters were not included in the mapping (Access: Re-implement Mantid [4] from binaries, Access: Re-implement Mantid from algorithms, Access: Emulation previous environment) as they relate to costs entailed in reconstructing software or runtime environments in order to gain access to past data which are specific to the ISIS facility and so do not seem to be covered within the standard. The ISIS cost model is not published and available for re-use as it is tailored specifically to the needs of the ISIS facility at STFC and therefore not intended to be generally applicable.
- KRDS – All parameters could be mapped successfully, however, some were based on certain assumptions due to the lack of definitions, which made it difficult to judge whether parameters corresponded directly to the ISO standard headings. Parameters that partly match the standard are mapped to the nearest relevant heading. Furthermore, KRDS comes with a benefit orientated toolkit as part of the cost model which might cover areas of the ISO standard that could not be covered by cost parameters such as Governance and Organizational Viability or Security Risk Management. The KRDS model covers a lot of areas of the standard but none of them overly extensively. Therefore, KRDS might be a good model to get a general overview without too much of a detailed insight. As mentioned in the benefits section of this report, the model provides a very useful benefits analysis toolkit.
- LIFE3 – There are a large number of cost parameters and most of them could be mapped against the detailed headings of the ISO standard. Some areas, however, are not clearly covered and might, for example, be covered within project management duties. This might be

dependent on the institution and the available input data. Further detailed headings under the section Governance and Organisational Viability that are not covered are: audit trail of changes to operations, procedures, software, hardware and self-assessment - external certification. In addition, two other areas seem to be not fully covered: Preservation Planning, although an automatic calculation is included and Security Risk Management, although disaster recovery is measured. Overall, the LIFE3 model maps well to the ISO standard and provides the best coverage of the models surveyed. Thus, it can be recommended for all areas of the standard.

- PrestoPRIME – Mapping parameters was very difficult as the cost model focuses on technical costs mostly in terms of costs per GB. Furthermore, the set of parameters is small and the parameters are not very well defined, if at all. However, all parameters were mapped, although some of them were based on assumptions or were partly matched to the nearest relevant heading. In addition, two parameters were mapped to high level headings of the ISO standard as it was not possible to map to the relevant sub-headings without having more information about the cost parameters. The PrestoPRIME cost model focuses on costs for hardware and software; staff costs are not included or only indirectly in terms of costs per GB. This model can be recommended for calculating the technically-based costs for archiving audio-visual material.

2.2.2 Cost parameters not mapped

- CET – Following a further review of this cost model and due to the level of detail provided within the Technical Guidelines which was extensive, it was decided that the CET model would not be investigated further. The parameters initially reviewed did not correspond to any headings of the ISO standard and therefore could not be mapped. However, for any future work related to the mapping exercise, the technical guideline documents should be considered. Furthermore, using the toolkit requires work on the input information at quite a granular level.
- OECD – Mapping proved to be difficult and not relevant to this exercise as the scope of the model is not digital preservation. The cost model provides transparent measures of administrative costs and its parameters could be mapped to headings in the area Organisational Infrastructure of the ISO standard based on a vague relation. The model could be used in conjunction with another cost model to determine administrative costs.

2.3 GAPS IDENTIFIED

The gaps identified by mapping cost model parameters to the ISO standard on audit and certification of trustworthy digital repositories are presented below by the ISO headings; Organisational Infrastructure, Digital Object Management and Infrastructure and Security Risk Management. The findings are based on the completed mappings of the eight cost models.

2.3.1 Organisational Infrastructure

As Table 2 shows, cost parameter coverage of the analysed cost models is not consistent within Organisational Infrastructure. When compared to the other areas, this shows the least coverage.

The KRDS model covers this area of the standard completely although two of the parameters in this section were mapped to the nearest relevant heading. On the other hand, the ENSURE and ISIS cost models do not cover any areas in this section. Governance, organisational viability consisting of mission statement, preservation strategic plan and collection policy, is only covered by a partial mapping of KRDS and PrestoPRIME parameters. LIFE3 also provides references to repository administration and management but as the parameters are not explicit, none of them were mapped to the headings above. Overall, with five of eight cost models being mapped, organisational structure and staffing as well as procedural accountability, preservation policy provide the best coverage of cost parameters across the cost models.

Table 2: Parameter mappings for Organisational Infrastructure

Organisational Infrastructure						
Cost Model	Sub-heading	Governance, organisational viability	Organisational structure and staffing	Procedural accountability, preservation policy	Financial sustainability	Contracts, licenses and liabilities
CMDP			X	X		
DANS			X			X
DP4lib			X	X	X	X
ENSURE						
ISIS						
KRDS		x	X	X	X	X
LIFE3			X	X	X	X
PrestoPRIME		x		X	X	

Key: x = cost parameter partially mapped
X = cost parameter fully mapped

The detailed headings provided within the ISO are listed below. The next level of analysis focused on the coverage of cost parameters at this level.

Governance and organisational viability

- *Mission Statement on the preservation of digital information*
- *Preservation Strategic Plan*
- *Collection Policy*

Organisational structure and staffing

- *Staff with adequate training and skills*

Procedural accountability, preservation policy

- *Defined Designated community and knowledge base*
- *Preservation Policy to ensure Preservation Strategic Plan met*
- *Audit trail of changes to operations, procedures, software, hardware*
- *Transparency and accountability in all actions supporting the operation and management of the repository that affect the preservation of digital content over time.*
- *Information integrity measures to be defined, collected and tracked*
- *Self-assessment, external certification*

Financial sustainability

- *Business planning processes*
- *Financial practices and procedures, 3rd party audits*
- *Analyze and report on financial risk, benefit, investment, and expenditure (assets, licenses, liabilities)*

Contracts, licenses and liabilities

- *Contracts or deposit agreements for digital materials*
- *Intellectual property rights management*

Looking at the analysis at this level, where details are provided in Annex 2, the following areas were only covered by one cost model; mission statement, preservation policy, self-assessment, external certification. As stated in our previous report, the reason that cost parameters cannot be provided for this area in general could be that measurable parameters are not able to be provided for this type of activity. It may also be the case that these areas are outside the scope of the cost models mapped. DANS and CMDP only cover two of these 15 detailed areas and therefore cannot be recommended for organisational infrastructure. LIFE3 provides the best coverage with 14 of the 15 detailed headings (some based on certain assumptions) and, therefore, is recommended to provide an overview of organisational infrastructure costs.

2.3.2 Digital Object Management

All eight cost models have some level of cost parameter coverage in the section of digital object management as shown in Table 3. Generally, this would be expected as cost models related to Digital Preservation should cover these areas.

ENSURE and LIFE3 provide the best coverage with parameters matching every heading of the ISO standard with two cost models, PrestoPRIME and DANS covering four out of six areas. The best coverage of cost parameters across these headings is for the two areas of the ingest process as they have matching cost parameters across all cost models.

Table 3: Parameter mappings for Digital Object Management

Digital Object Management							
Cost Model	Sub-heading	Ingest: acquisition of content	Ingest: creation of AIP	Preservation planning	AIP preservation	Information management	Access management
CMDP		X	X	X	X		x
DANS		X	X	X		X	
DP4lib		X	X		X	X	X
ENSURE		X	X	X	X	X	X
ISIS		X	X	X	X	X	
KRDS		X	x	X		X	X
LIFE3		X	X	X	X	X	X
PrestoPRIME		x	x			x	X

Key: x = cost parameter partially mapped
X = cost parameter fully mapped

The detailed headings provided within the ISO are listed below. The next level of analysis focused on the coverage of cost parameters at this level.

Ingest: acquisition of content

- *Content Information and the Information Properties*
- *Content Information on deposit*
- *Specifications enabling recognition and parsing of the SIPs.*
- *Verification of Producer identity*
- *Ingest process to verify SIP*
- *Preservation control over the Digital Objects*
- *Responses to producer/depositor*
- *Content acquisition records*

Ingest: creation of AIP

- *Definition for parsing AIP and for long term preservation*
- *AIP construction form SIP*
- *Document SIPs*
- *Persistent unique identifiers for AIPs*
- *Provide Representation Information (RI) for all digital objects*
- *Preservation Description Information (PDI) for Content Information (CI)*
- *IP Content Information understandable to Designated Community*
- *AIP verification*
- *Repository integrity verification*
- *AIP creation records*

Preservation planning

- *Preservation strategy*
- *Monitor preservation environment*
- *Preservation plan changes*
- *Preservation activity effectiveness*

AIP preservation

- *AIP storage spec*
- *AIP storage and preservation records*

Information management

- *Discovery and identification of material by Designated Community*
- *AIP descriptive info*
- *Bi-directional linkage between AIP and descriptive info*

Access management

- *Access policies*
- *Authenticity of Digital Objects*

It was evident, as provided in Annex 2, that even at this level of detail the standard is covered quite well with at least three of the eight cost models covering every section, except for the area of preservation activity effectiveness. The PrestoPRIME and ISIS models covered the least headings in this section. Overall, LIFE3 provides the best detailed coverage with 24 out of the 29 detailed headings mapped.

2.3.3 Infrastructure and Security Risk Management

The section on infrastructure and security risk management has good coverage as every cost model matches at least one of the two sub-headings as shown in Table 4.

CMDP, PrestoPRIME and LIFE3 cover this section best as they have parameters mapped to both areas. Technical infrastructure risk management is covered by seven of the eight cost models with security risk management being only covered by half of the cost models mapped.

Table 4: Parameter mappings for Infrastructure and Security Risk Management

Infrastructure and security risk management			
Cost Model	Sub-heading	Technical infrastructure risk management	Security risk management
CMDP		X	X
DANS			X
DP4lib		X	
ENSURE		X	
ISIS		X	
KRDS		X	
LIFE3		X	X
PrestoPRIME		X	X

Key: X = cost parameter fully mapped

The detailed headings provided within the ISO are listed below. The next level of analysis focused on the coverage of cost parameters at this level.

Technical infrastructure risk management

- *Identify and manage risk management to preservation operations and system infrastructure*
- *Manage number and location of copies of Digital Objects*

Security risk management

- *Analysis of security risk factors associated with data, systems, personnel, and physical plant*
- *Security risk controls*
- *Staff roles and responsibilities*
- *Disaster recovery plans, one off-site backup and offsite recovery plan*

As provided in Annex 2, the six detailed headings in the section are not covered consistently. Less than half of the headings in this section were covered. ISIS and DP4lib cost parameters could only be mapped to one detailed heading. Staff roles and responsibilities under security risk management were covered by only one cost model, namely DANS. Two out of eight cost models matched parameters against "analysis of security risk factors associated with data, systems, personnel and physical plant" and "security risk control". With parameters matching five of the six detailed headings, the CMDP model provides the best coverage for this section.

Summary of gaps identified Table 5 provides an overview of the gaps identified by mapping the parameters of cost models for digital preservation against the ISO 16363 standard on audit and certification of trustworthy digital repositories. This exercise clearly shows that the gaps identified relate to governance and organisational viability as this is covered by only two out of eight cost models.

At a more detailed level of analysis, gaps are: mission statement, preservation policy, self-assessment and external certification, all in the section: organisational infrastructure, and staff roles and responsibilities, part of the section Infrastructure and security risk management.

There may be very legitimate reasons for these gaps as they could be out of the scope of the cost model, or they are not realistically measurable as a cost parameter. Institutions that have created a cost model would have matched their activities and workflows to the parameters defined, for example, preservation and collection policy might exist in some but not others. Staff roles and responsibilities is a similar heading that is measured very precisely by some cost models and influences several other cost parameters in activity based cost models.

Looking at the different approaches the cost models take, it is not possible to exclude some of the ISO standard headings completely from the mapping. However, one should keep in mind that some areas of the standard are not covered by measurable cost parameters and covering all parts of the standard is probably not possible with quantitative cost parameters, but by looking at qualitative aspects these areas could be covered.

Table 5: Summary of gaps identified when mapping cost models to ISO16363

ISO heading	Organisational Infrastructure					Digital Object Management						Infrastructure and security risk management	
Sub-Heading / Cost Model	Governance, organisational viability	Organisational structure and staffing	Procedural accountability, preservation policy	Financial sustainability	Contracts, licenses and liabilities	Ingest: acquisition of content	Ingest: creation of AIP	Preservation planning	AIP preservation	Information management	Access management	Technical infrastructure risk management	Security risk management
CMDP		X	x			X	X	X	X		x	X	X
DANS		X			X	X	X	X		X			X
DP4lib		X	X	X	X	X	X		X	X	X	X	
ENSURE						X	X	X	X	X	X	X	
ISIS						X	X	X	X	X		X	
KRDS	x	X	X	X	x	X	x	X		X	X	X	
LIFE3		X	X	X	X	X	X	X	X	X	X	X	X
PrestoPRIME	x		X	x		x	x			x	X	X	X

Key: x = cost parameter partially mapped
X = cost parameter fully mapped

2.4 AREAS FOR FURTHER INVESTIGATION AND DEVELOPMENT

Following the first report, D32.1, the analysis of cost parameters concentrated on finalising the mapping exercise and completing the gap analysis to a detailed level. As mentioned previously, gaps were analysed where cost models and their cost parameters did not cover the headings of the ISO standard. The areas identified as true gaps evidenced by the mapping exercise provide cost model owners and developers with a good starting point for reviewing cost models. This would of course, only be relevant if the cost model's purpose is to cover all areas of the standard on audit and certification of a trustworthy digital repository. Given that this was our starting point and provided a sound basis for comparison it may be that we consider that all of our digital repositories are able to be costed through the use of a cost model in this way. There are areas which may require additional focus for cost model owners and the analysis and mappings provide recommendations for specific areas.

The gaps identified should be considered in conjunction with the respective scope and level of detail which depends upon the purpose of the organisation for which the cost model was generated. The main gaps identified were in the organisational infrastructure section related specifically to governance and organisational viability. If the cost model is only to be used to estimate the annual budget, it may not be necessary to cover the governance issues of the organisation. However, each cost model should at least cover the aspect of organisational viability where practicable. In the case of the DP4lib model, the gaps identified provide an increasing emphasis on the section of organisational viability.

The main area which should be looked at for further development which would allow for cost models to become more useful to a wider audience is cost parameter definitions which need to be provided in a clear, concise and understandable form and have been found to be lacking in a number of cases.

2.5 SYNTHESIS OF COST MODELS – SIMILARITIES AND DIFFERENCES

With the mappings completed and the detailed analyses, we could draw some conclusions in relation to the similarities and differences between the models. In terms of similarities, the mapping exercise showed that there are several areas which are covered quite well by several cost models. Examples are the areas of ingest (ingest: acquisition of content, ingest: creation of AIP) and risk management (under the detailed heading of, identify and manage risk management to preservation operations and system infrastructure).

A similarity that was evidenced in the general analysis of the cost models was the difficulty to re-use them. Most of the cost models can be mapped to the ISO standard, however, they are usually tailored to the cost model owner's needs. The problems of cost model re-use are illustrated in detail in the section on testing of cost models in this report.

When it comes to differences, it can be stated that the cost models vary in breadth of cost parameters and therefore the degree in which they cover the ISO standard. The ISIS cost model for example has limited mappings of cost parameters to the standard, whereas, the LIFE3 model mapped well across most areas. Therefore, comparing cost models is very difficult given the level of granularity provided by the descriptions of cost parameters which varies significantly from one cost model to another.

3 TESTING OF COST MODELS

3.1 INTRODUCTION TO TESTING UNDERTAKEN

Testing of a cost model is an important step towards its validation. The more a cost model is tailored to the individual business processes of an institution, the harder it is to test it with data from a different institution. Most cost models investigated in APARSEN were tailored to institutions' specific needs. This is understandable from the cost model creator's point of view because the tailoring makes the cost model more useful in the context of the individual's institution. It raises the question, however, of how different cost models can be compared with each other and more importantly, be used by a wider audience. The answer is arrived at through testing but there is another kind of validation and testing by case study, although not in scope here, which would look at how well the model worked in the organisation it was created for and whether costs have been accurately modelled over a period of time. This would allow for both the validation of the model and give a sense of how a given model related to a given service context or set of business practices.

Three cost models: DANS, DP4lib and LIFE3 were selected based on the criteria as provided in section 1.1.2 of this report. Two sources of test data were used in this exercise, from DANS and DP4lib. The testing of the DANS model with DP4lib, and vice versa provided in-depth analyses across these two cost models, the detailed results of which are provided in section 3.3 below.

There is an argument that to be objective at least one independent source of data should have been tested across all three models, however, in practice, as previously stated this was not possible.

Table 6: Cost models and test data sources

Cost Model	Test data source
DANS	DP4lib
DP4lib	DANS
LIFE3	DP4lib

3.2 TESTING THE DANS COST MODEL

3.2.1 Overview of the DANS model

The DANS cost model is an activity-based model for the long-term preservation and dissemination of digital research data. The model calculates the costs of archiving datasets and is based on activity-based costing and the balanced scorecard. The model covers research data archives and is under development with validation still to be finalised. The largest digital collections are in archaeology, social sciences and history. For these disciplines DANS offers services in digital curation and long term preservation. Until now these services have been provided without a fee, but in the near future this will change, first for archaeology and later for the other disciplines as well. For this reason DANS has been developing its cost model to get a clear understanding of all costs involved in providing this service. The model was developed by DANS, Data Archiving and Network Services, the research data archive of the Netherlands, funded by the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Netherlands Organisation for Scientific Research (NWO), the National Research Council of the Netherlands.

The DANS cost model contains a major distinction between direct and indirect costs.

<u>Direct costs</u> are costs directly related to the data archiving process, mainly labour costs of archivists, acquisition and licence costs. Assigning cost drivers to the direct costs is relatively easy and straightforward.

<u>Indirect costs</u> are all other costs incurred by the repository, ranging from office rental payments, through labour costs of non-archivist personnel to storage costs. The indirect costs are distributed amongst all the activities. This is a very complex operation. All activities in the repository are classified as either principal tasks or auxiliary tasks. Principal tasks are considered to be all activities

to do with the core business of the archive or repository, and auxiliary activities as those which are also needed (like office management) but which are not core business activities.

All costs of the *auxiliary activities* are distributed into the *principal activities* by first splitting them up in cost centres for which cost drivers are defined. This is required because the distribution of costs of the auxiliary activities cannot be allocated to the costs of the principal activities in the same proportion. For example the auxiliary activity costs related to “communication” should only be allocated to the principal activity “preparation of the archive newsletter” and not to other principal activities like “ingest” or “conversion of original files”. This is a rather arbitrary distribution scheme. What is finally obtained are the *total indirect costs* for each principal activity, consisting of the auxiliary costs and the costs of the principal activity.

For 2010, direct costs of data archiving and labour costs accounted for 44% of total costs, whereas indirect costs of office, IT support, other labour costs, material costs, KNAW services, etc. accounted for 56% of total costs.

The outcome of the cost model is of course strongly dependent on how direct costs are defined. In the model these are now mainly labour costs of the archivists in the strict sense of the word. There are, however, other costs, now defined as indirect costs, which could be just as easily classified as direct costs as well. This could be the case in particular with storage costs.

The consequence of this method is that both direct and indirect costs can be compared easily between possible domains of the repository. Also the proportion of direct costs within the total costs can easily be seen and compared between the domains. For the DANS organisation this has been done for three main domains of the repository: archaeology (small and large), social sciences and history.

3.2.2 Testing the DANS model with DP4lib test data

Testing the DANS cost model with data from the DP4lib cost model was a complex operation. In the DP4lib model all cost elements are broken down into five different cost categories: hardware, software, staff, accommodation and external services. The costs in these categories are certainly included in the DANS model; they are however not broken down in the same way.

With allocation and apportionment based on assumptions a comparison on a high level has been carried out.

<u>Direct costs</u>

Licence costs could be taken relatively easily from the DP4lib model and categorised as direct costs. For the labour costs of archivists and acquisition officers, in practice this was not always distinguishable in the DANS organisation; as a result the labour costs of the assistant and executive officer have been chosen as the comparable costs from the DP4lib data.

<u>Indirect costs</u>

Here in fact all the other costs within the DP4lib model could be accommodated. It was not possible to follow the subdivision into principal activities as followed in the DANS cost model, because these are composed in a totally different way. So the apportionment of the DP4lib data had to be followed here under the main activities; hardware, staff (in this case other than the staff members mentioned under direct costs), accommodation and external services.

As the DP4lib data is based on another type of organisation, the allocation and apportionment of costs could not be made into the subdivisions of the different domains within the DANS model. However, the possibility to do this remains within the DANS cost model, if these figures were made available.

All the costs have been apportioned under the main activities; Ingest, Curation and Access according to the same percentage allocations as in the DP4lib model.

Only a comparison on a high level could be carried out. Theoretically, for future testing, it would be possible to test at a more detailed level, if all the costs of the DP4lib model were available at this level. To carry out the division of auxiliary tasks over the principal tasks by using cost centres would then be possible. This is quite a labour-intensive task as the DANS cost model consists of at least ten worksheets in which the grouping, calculation and subsequent apportionment and allocation of the principal and auxiliary tasks is performed.

3.2.3 Findings

As has been concluded the overall division of direct and indirect costs is very different within the DP4lib and the DANS cost models. In the DANS cost model filled with DANS data 44% of all costs are direct and 56% are indirect; in the DANS model filled with DP4lib data only 23% are direct costs and 77% are indirect costs, as shown in Figure 3. The comparatively high percentage of direct costs in the DANS model is mainly due to labour costs. The differences between DANS and DP4lib may be due to intrinsic differences in their services or activities and the costs involved, or because the cost categories and allocations used by DANS are not fully incompatible with the DP4lib data.

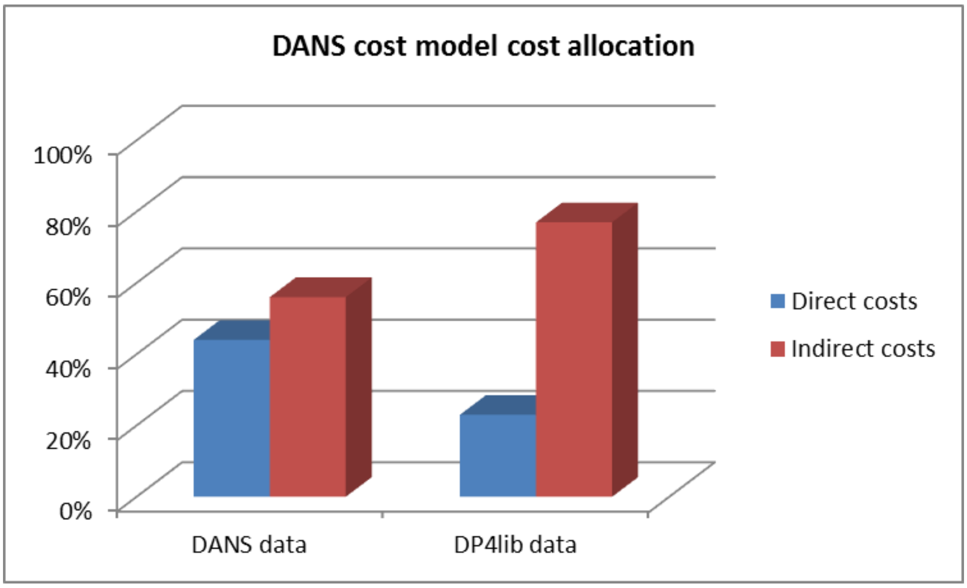


Figure 3: Allocation of costs in the DANS model

3.2.4 Conclusion

On a very high level a comparison between the two models was possible. At a more detailed level it would only be possible by using all the detailed figures which was not feasible within this project due to limitations around access to the data.

Generally, when using another cost model, even at a sufficiently high level, it is necessary to make explicit all the underlying basic assumptions of the model being tested as well as the test data being utilised. If these assumptions are not available, then even a comparison at a high level is a difficult if not potentially flawed exercise. When these assumptions are known, however, testing cost data in another model could give valuable insights into both models as well as to the contexts in which these cost models were created.

3.3 TESTING THE DP4LIB COST MODEL

3.3.1 Overview of the DP4lib model

The German National Library (DNB - Deutsche Nationalbibliothek) is working, together with other partners, in the “Digital Preservation for libraries” (DP4lib) project to establish a replicable and flexible infrastructure for long term preservation. Software as well as concepts are being developed in a generic way where possible. This also applies to the selection of a business model covering the whole preservation process. All elements of the resulting prototype should be usable for every heritage institute tasked with long term preservation of digital material. This means that the DP4lib cost model could be used by digital repositories in libraries, archives or other heritage institutes.

Within the project a service portfolio was generated, which consists of more than two dozen large and small services for long-term preservation and controlling activities. To be able to create a cost model for the operation of a long-term preservation (LTP) service the DP4lib portfolio was distributed into three main areas: Ingest, Curation and Access. The OAIS functional model has been used to define these areas as shown in Figure 4.

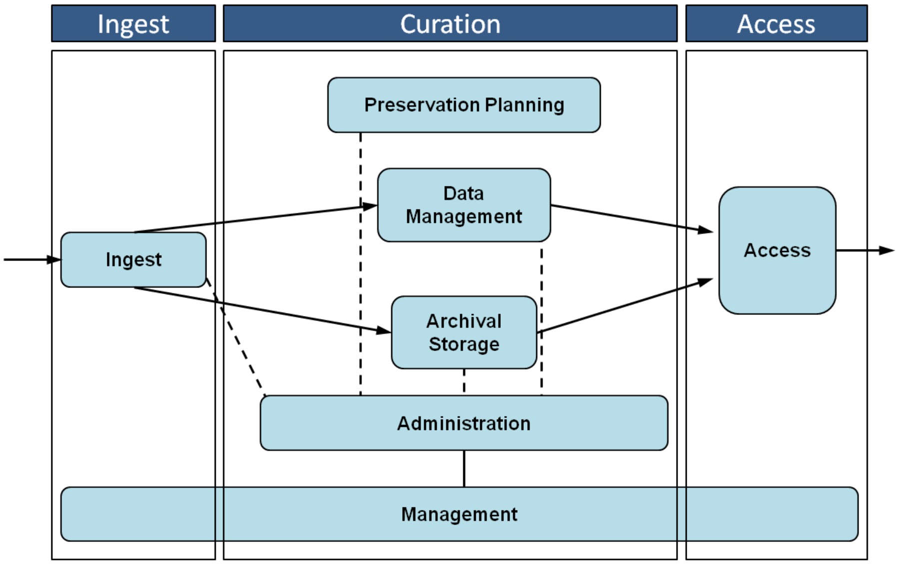


Figure 4: Distribution of services in the OAIS functional model

Table 7 below shows the main services with their respective sub-services as defined within the DP4lib model:

Table 7: Services and sub-services of DP4lib model

Ingest	Curation	Access
Reception of the objects	Digital Lifecycle Management	Authentication
Metadata handling	Conservation activities	Search
SIP handling	Integrity check and conservation	Retrieval
Reporting and Protocol Management	Retrieval (Search and Access)	
Storage of Objects		

For each of these three main services a Cost-by-Service cost model was established. The result of the considerations made on this basis, is a spread sheet which is designed to support service providers in calculating the costs of the DP4lib-LTP service.

3.3.2 Testing the DP4lib model with DANS test data

Testing the DP4lib model with cost data from DANS was a complex exercise due to structural differences between the DP4lib and DANS cost models. In the DP4lib model, for example, all cost elements were broken down into five different cost categories: hardware, software, staff, accommodation and external services. This breakdown allows for general insights into the costs of digital preservation, for example, the annual cost for hardware at the DNB are, according to the DP4lib model, for all cost elements, €146,000. If the breakdown of these costs was the same within the DANS cost model, general comparisons could be possible with regard to hardware costs. Since the costs could not be made comparable in these ways, a higher level of comparison had to be found.

Initially, some common ground between the DP4lib and the DANS models was sought in spite of very different digital preservation workflows and IT infrastructures. Comparing all cost elements of the two models, three matching categories are found: staff, external services and software. The DANS cost model, however, also distinguishes between direct and indirect costs. Whilst direct costs are always assigned to one of the three mentioned categories, indirect costs are assigned to 20 main and 15 secondary processes that support the overall digital preservation services. For this exercise, cost data for four digital preservation services were available: Archaeology (small), Archaeology (large), Social science and History.

All components of the main and secondary processes are broken down in detail and distributed over the four preservation services. However, they include staff costs due to their 'indirect' nature, which are listed here and not in the main cost category for staff. In the DP4lib model, such distinctions are not made as all staff-related costs appear only in the respective main category. Therefore, in order to allow for a better comparison of the two models, a new category of indirect costs was defined for matching purposes. All indirect cost elements were subsumed under the categories defined for each cost model i.e. the categories of hardware and accommodation from the DP4lib model and all 35 main and secondary processes representing indirect costs in the DANS model.

The next stage was to test the transfer of the DANS cost data from the DANS cost model to the DP4lib cost model. Because both models are based on activity-based costing, this transfer was conceptually possible. As mentioned previously, the DANS model distinguishes between four services and between direct and indirect costs which are distributed according to these services. The DP4lib model distinguishes between three main processes, with 28 different activities (or sub-processes) distributed according to the areas of; ingest, curation and access.

The level of detail in cost breakdown and distribution between the DANS and the DP4lib models differ significantly. For example, in the DP4lib model, the annual cost of a server is distributed between ingest, curation, and access according to the CPU time for the respective computing activities. Because the DANS model uses different distribution keys, a direct transfer of the data and a detailed comparison is not possible.

For one of the services, archaeology (small), an attempt was made to distribute the 15 relevant principle activities within the indirect costs over the activities of ingest, curation and access from the DP4lib model. However, for indirect costs an approximation could only be made which allowed for an approximately fair distribution of direct costs as not enough information was available. Figure 5 shows the respective proportions of ingest, curation and access activities within the archaeology (small) service.

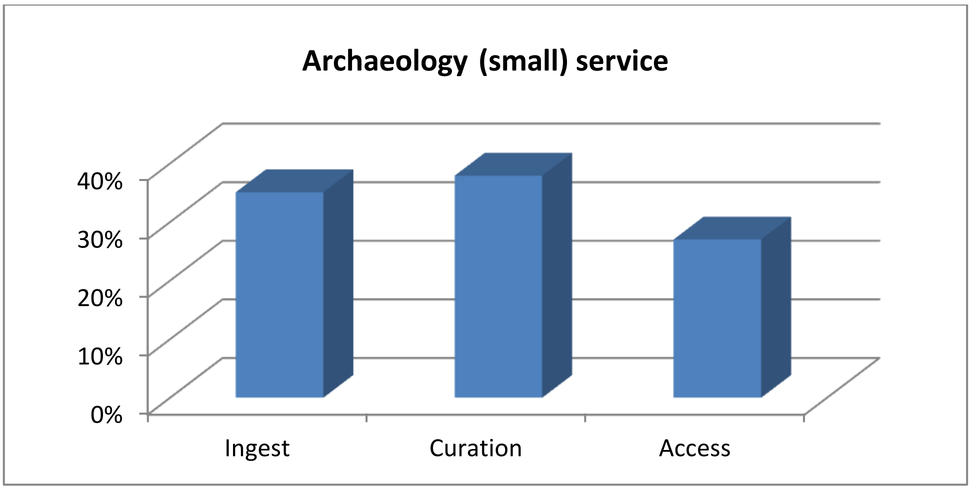
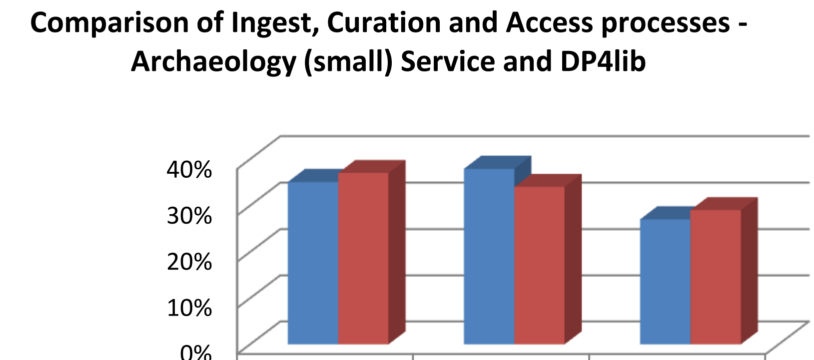


Figure 5: Distribution of the total costs of a DANS service by DP4lib processes

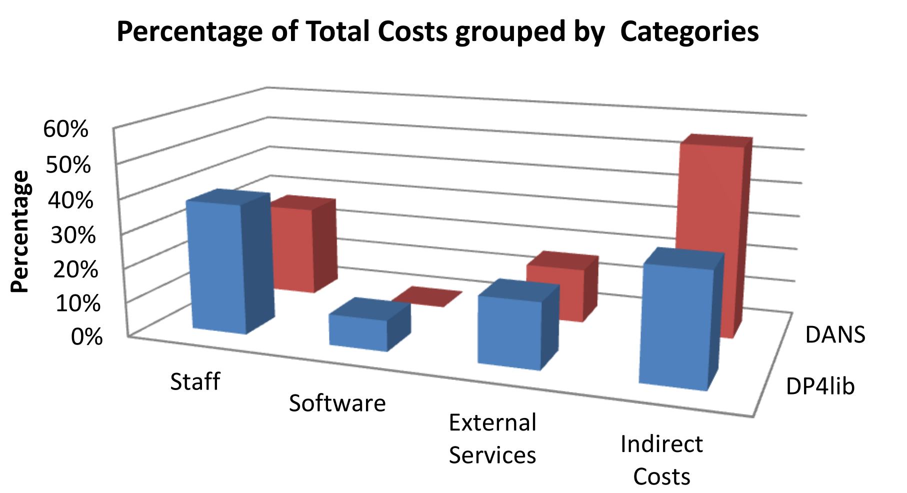
It is worth noting that by comparing the DANS data in relation to the DP4lib processes (ingest, curation, access), the percentage of total costs for the respective processes are similar across both models. When indirect costs are considered for all ingest activities for the Archaeology (small) service this makes a contribution of 35% of the total costs of the service, whilst this process represents 37% of all costs in the DP4lib model as shown in Figure 6. Similar trends can be found in the other two categories.



	Ingest	Curation	Access
Archaeology small	35%	38%	27%
DP4lib	37%	34%	29%

Figure 6: Comparison of processes between a DANS service and the DP4lib model

The next analysis undertaken was of the categories used across the two models which included indirect costs and is shown in Figure 7.



	Staff	Software	External Services	Indirect Costs
DP4lib	38%	9%	19%	32%
DANS	27%	0,13%	16%	55%

Figure 7: Comparison of categories across the DP4lib and DANS models

The apportionment of external services costs seemed to be straight-forward. The DANS test data provides clear assignments as does the DP4lib model so this category can be regarded as a good test of the transferability of these different costs between the two models. However, the transfer of software costs proved more difficult as within DP4lib only the license costs related to commercial software were included. However, transferring the DANS test data into this category shows that only 0.13% of all costs can be included in this category. All other software costs identified in the DANS test data could not be assigned so were included under indirect costs.

A very good transfer of DANS test data was found for staff costs as all direct costs focusing on personnel costs were grouped into the category of staff in the DP4lib model. However, this applies only for the direct costs mentioned in the DANS test data. All indirect personnel costs, which are included in the 35 activities of the DANS model, could not be analysed in more detail. Due to these problems indirect personnel costs had to be left in the category of indirect costs; on the other hand a detailed breakdown of direct personnel costs was possible. Figure 8 shows the different apportionment of staff cost depending on the four DANS services.

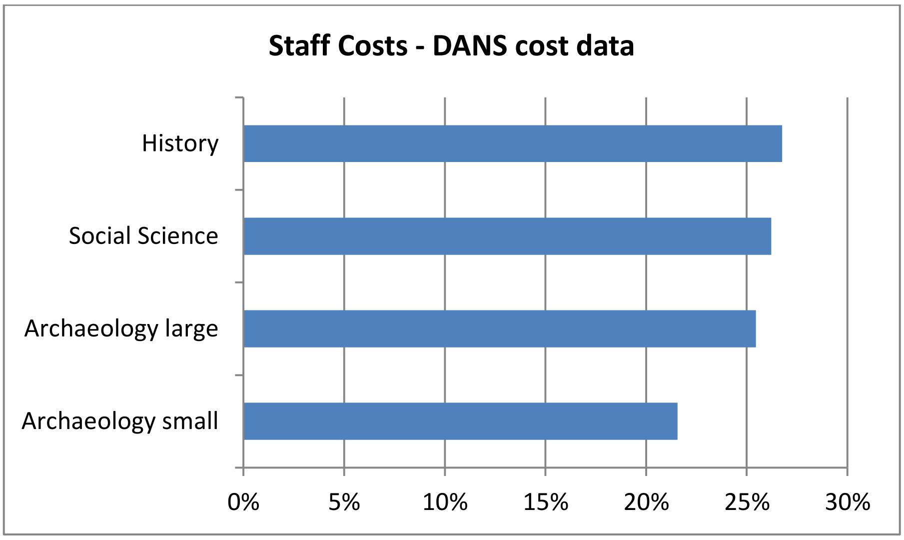


Figure 8: Staff costs for DANS services

In DP4lib two additional categories of hardware and accommodation were used; however, no correlations could be found in the DANS test data, so further analysis of the DANS test data could not be carried out for these costs.

3.3.3 Findings

In testing the DP4lib model with the DANS test data, only 45% of all costs were able to be allocated. About half of the DANS data had to be sorted into the new category of indirect costs and were therefore not allocated. This was the reason why this cost category as shown in Figure 7 was the highest.

Since it is generally possible to set up a new category in the DP4lib model in order to include other cost categories (indirect costs), it is possible to perform 100% apportionment of all DANS data into the DP4lib model at a high level.

3.3.4 Conclusion

Conceptually and on a very high level, a comparison of the DP4lib cost model and the DANS cost model is feasible. Both models are based on activity based costing so that the data can – up to a certain level – be allocated between the two models. However, the creation of a new, normalized costing category for various, otherwise incomparable, cost elements was necessary in order to complete the exercise.

The comparison on the basis of the analysed processes and on the basis of the categories allows some conclusions with regard to annual staff cost or the extent of indirect costs. Because of the very different implementation and modelling of digital preservation activities, as well as the different level of detail and distribution methods, a more detailed analysis of both models is not feasible.

For any future testing or validation it is recommended that data from within the organisation, where the same processes have been implemented, be used for this exercise. Where external data is used a clear understanding of the test data is required including any assumptions made in relation to cost allocations. It may be a useful exercise to look at the data being used by another organisation within the context of these models as lessons can be learned about how costs are measured, distributed and allocated.

3.4 TESTING THE LIFE3 COST MODEL

3.4.1 Overview of the LIFE3 model

The LIFE3 model provides a predictive costing tool for the long-term costs of digital preservation (DP) for DP repositories. The model uses case studies to predict costs and provides the values of specific parameters. Because of lack of historical figures, estimates of costs have been made. The model represents costs related to preservation activity over time. Trends in areas such as tool development and the life expectancy of file formats have been estimated and modelled. Inputs to the model for base costs like staffing have been identified and defined. As the model is pre-populated an estimation of costs for long term preservation is provided relatively easily with the input of basic data. For more accurate figures the model allows for data to be updated to the user's needs.

3.4.2 Testing the LIFE3 model with DP4lib test data

It was important to understand and compare the two models. Firstly, as the LIFE3 cost model owner was not a project partner, it was important to carry out some research into the model by referring to user guides (where available) or relevant documents published. Secondly, the test data had to be reviewed and understood in terms of cost breakdowns and allocations by working with the owner of the data and their respective cost model.

The DP4lib model is a cost-by-service cost model. Costs are costs are allocated along the service lines of ingest, curation and access. The total cost calculated provides the cost allocation on an annual basis in terms of storage capacity of the system in GB and provides the amount invoiced to the customer. The cost model supports a non-profit centre. Therefore, the cost to the customer does not include any element of profit. Overheads are accounted for under the 'Accommodation' category. The model was created with the help of an economist. The test data is based on estimates and some of these may require further assessment and may need to be clarified by measuring methods. Test data provided from the DP4lib model gives total annual cost of € 622,053 for a storage capacity of 100,000GB, which is broken down into processes as shown in Table 8 below:

Table 8: Costs of processes in the DP4lib model

	Cost of process (€)	Cost of process per GB (€)	% of total cost
INGEST	230, 981	2.31	37%
CURATION	212, 665	2.13	34%
ACCESS	178,407	1.78	29%

The costs under these processes are broken down further into the cost categories of hardware, software, staff, accommodation and external services. These are then broken down into 21 cost elements and distributed or apportioned (as % of annual cost) across the three processes using allocation methods, namely, distribution keys. These distribution keys, depending on the type of cost, are: CPU time, process, storage, working time and service hours.

The aim of the exercise was to ensure that all DP4lib costs in the form of test data could be entered into the LIFE3 model. An iterative approach was undertaken providing – basic input, further detailed entry of data (where possible) and a final allocation of costs at high level.

At the basic level of entry, the LIFE3 model allows for entry of data into the basic input work sheet at a high level with minimum input to provide an initial predicted lifecycle cost.

The data entered was:

- Timeframe one year – 2013, as the test data provided annual costs
- Category - e-journals were selected for the purposes of testing the DP4lib data. The DP4lib model makes no such distinction of file formats with a diverse pool of publications, excluding web sites and research data.
- Source – provides the method by which the material to be preserved will enter the digital repository. Donated was selected. The DP4lib service provides a long term preservation service which would require no processing prior to ingest as assumed with the costings of the model, so this is considered free of cost, which relates to the definition of donated items from the LIFE3 model.

- Number of files to be selected for the year
  The definition given is: "for each year of the modelling period selected, the number of items to be selected needs to be entered. Selection means that the items will be ingested into the repository, and if necessary digitized or harvested beforehand. The input fields should be left blank for years in which no new material is to be selected, with existing material being preserved only."
  The unit of measure is important here as the DP4lib test data provides storage capacity in GB whereas the LIFE 3 model allows entry of number of pages, sites, e-journals etc. As the DP4lib cost model provides cost for 100,000 GB of storage capacity for their long term preservation service, the number of files selected should equate to the same storage size. The LIFE3 model provides storage sizes by file format as shown in the table below:

LIFE3 options available for selecting material	LIFE3 – average storage size in MB
Web sites	256MB per web site
E-journals	**1MB per e-journal**
Printed items	30MB per digitized print item
Sound recordings	500MB per sound recording
Other	1MB per item
Research documents (up to 10Mb)	0.3MB per research document

  For testing, a conversion was made from average storage size to items selected as the unit of measure required for testing 100,000GB of storage capacity which related to the DP4lib test data. This was equivalent to 102,400,000 e-journals (as 1GB = 1024MB and 100,000GB = 102,400,000 MB = 102,400,000 e-journals).
- Organization type – medium was selected as this was the nearest match to the DP4lib organization.

The model can be customized further as the composition of the repository architecture can be changed in detail later, by refining the Organization Profile and Bit-stream Preservation data. This was not done in this exercise.

On calculating the cost, an output work sheet is generated, see Figure 9. The costs were selected at net present value so no discounting was applied to the figures.

Estimated Lifecycle Cost

Please note that the figures provided on this page are estimates only. You can increase confidence in these figures by editing the default values the model is using. To do so, please click on the tabs beginning with 'Refine…' at the bottom of the screen, then review and where necessary change the values being used for your organization and each of the model stages.

Lifecycle costs for 2013 to 2013

Calculate costs as Net Present Value? Y

Creation or Purchase	Acquisition	Ingest	Bit-stream Preservation	Content Preservation	Access
Donated	Selection	Quality Assurance	Repository Administration	Preservation Watch	Access Provision
£0	£0	£0	£3,600	£2,000	£0
	Submission Agreement	Metadata	Storage Provision	Preservation Planning	Access Control
	£0	£0	£160,000	£2,000	£0
	IPR & Licensing	Deposit	Refreshment	Preservation Action	User Support
	£0	£0	£0	£0	£0
	Ordering & Invoicing	Holdings Update	Backup	Re-ingest	
	£0	£0	£0	£0	
	Obtaining	Reference Linking	Inspection	Disposal	
	£0	£0	£47,000	£0	
	Check-in				
	£0				
£0	£0	£0	£210,000	£4,000	£0
0%	0%	0%	98%	2%	0%

Total Estimated Lifecycle Cost: **£220,000**

All figures on this sheet are rounded to 2 significant figures.

Figure 9: LIFE3 model test data at first level of entry

As shown in Figure 9, on converting to € (using an exchange rate of £ 1 = € 1.18) this provided an estimated lifecycle cost of € 259,600, giving an allocation of 42% of the total costs from the DP4lib model.

At this stage it was not clear whether the cost breakdowns were mapped to the correct lifecycle processes and sub-processes of the LIFE3 model (referred to as elements and sub-elements within the LIFE3 model) as this is an automated process. As the costs provided are estimates, these can be refined further to provide a more accurate costing by entering detailed cost information for organisation profile as well as the lifecycle processes of: creation, acquisition, ingest, bit-stream preservation, content preservation and access. The model variables are provided on further work sheets which show the costs pre-populated in the model through case study and set up data. Therefore, initially there may not be an accurate match to the relevant processes and sub-processes of your digital repository.

The next stages of the exercise were to enter data at a detailed level and those costs which had not already been allocated (i.e. 58%) from the DP4lib model by going through each work sheet to refine the data entered where relevant.

Refine organisation – the default staff data (salaries and roles) was overridden with costs from the DP4lib model as in Table 9 below:

Table 9: LIFE3 staff costs updated for DP4lib data

LIFE 3 staff costs		DP4lib staff costs	
Senior Manager	100,337 €	Manager	58,800.00 €
Project Manager	64,369 €	Executive officer	45,600.00 €
Technical Officer	51,429 €	Developer	45,600.00 €
Project Team	41,295 €	Techn. support	45,600.00 €
Operational staff	32,994 €	Assistant	40,800.00 €

Refine creation - is not relevant to the DP4lib services test data as all items to be preserved will be provided by the customer using the preservation service. Therefore, stating items are donated provides a zero cost – as required

Refine acquistion – no data was updated as the breakdown of processes did not match well against the costs within the DP4lib model. A review was undertaken to check if the any of the processes were in fact included within the ingest costs; it was found that only 'time spent on checking-in' matched

Refine ingest – the LIFE3 model allows only for the allocation of staff costs for this process. Staff costs were entered from the DP4lib model; however, the DP4lib model also allocates costs related to hardware, software, accommodation and external services to this process whereas the LIFE3 model does not allow for the entry of these other costs

Refine bit stream preservation – some of the curation processes (e.g. repository management) within the DP4lib model can be matched with entry as days of effort per year for staff at a particular grade. Data is already provided which may be an automatic calculation based on the data entered or pre-populated. However, data cannot be entered as provided by the DP4lib model, due to the unit of measure being different. Back up costs are entered

Refine content preservation - collection heterogeneity is set as low as e-journals are considered to be uniform. Emulation costs are entered as they are included within migration and can be calculated

Refine access - measured as days of effort per year for a staff member of a particular grade. No data was entered as this information is not available from the DP4lib cost data

The results of this exercise provided total costs allocated from the DP4lib model into the LIFE3 model of € 318,600, giving an allocation of 51%.

Subsequently, a review was carried out with the owner of the DP4lib cost model and it was decided (having learnt from other testing carried out, and the fact that the LIFE3 model provides for a very detailed level of input) that an allocation of costs at a higher level may prove to be more productive in allowing for further allocations of costs from the DP4lib test data to the areas of the output worksheet of the LIFE3 model. The methodology adopted for this exercise, for example, on analysing access costs, is detailed as follows: Access provision costs of €52,204 have been allocated from preparation costs consisting of staff (€18,408), hardware (€20,048) and external services (€13,748); Access control costs of €20,071 are allocated from authentication costs of staff (€5,522), hardware (€801) and external services (€13,748); and user support costs allocated as staff (€5,522), hardware (€3,207) and external services (€13,748).

The allocation method used was to create distribution keys. For example: Access – Process, looking at the "Preparation" sub-process, we have implemented a monitoring system that is able to observe the distribution keys of used CPU time, used disk cache and used storage for all cost elements collected in the hardware category. In the DP4lib model, all "Preparation" actions, which are all running on a dedicated server, were measured then by these three distribution keys for a month and these results were extrapolated to obtain the annual costs.

The resulting distribution values were:

Used CPU Time : 8% of all used CPU Time of the dedicated Server

Used Disk Cache: 40% of all used Disk Cache of the dedicated server

Used Storage (long time): 0%

Accordingly, we know that we must calculate the preparation costs as 8% of the annual cost of the dedicated server. For all other distribution keys, the same procedure was used.

For access costs variable and fixed costs are also calculated. Variable costs are incurred only if the process is running. Costs that are incurred even if the process is not running are summarized under fixed costs.

3.4.3 Findings

The results of this final test are given in Table 10 below:

Table 10: Final allocation of costs from the DP4lib model to LIFE3

Estimated Lifecycle Cost

Please note that the figures provided on this page are estimates only. You can increase confidence in these figures by editing the default values the model is using. To do so, please click on the tabs beginning with 'Refine...' at the bottom of the screen, then review and where necessary change the values being used for your organization and each of the model stages.

Lifecycle costs for 2013 to 2013

Calculate costs as Net Present VY

	Creation or Purchase	Acquisition	Ingest	Bit-stream Preservation	Content Preservation	Access
	Donated	Selection	Quality Assurance	Repository Administration	Preservation Watch	Access Provision
	£0	£0	£19,755	£22,780	£11,637	£44,241
		Submission Agreement	Metadata	Storage Provision	Preservation Planning	Access Control
		£0	£19,193	£29,473	£11,637	£17,010
		IPR & Licensing	Deposit	Refreshment	Preservation Action	User Support
		£0	£18,149	£0	£25,170	£19,049
		Ordering & Invoicing	Holdings Update	Backup	Re-ingest	
		£0	£0	£29,473	£0	
		Obtaining	Reference Linking	Inspection	Disposal	
		£0	£0	£29,473	£0	
		Check-in	only staff and hardware costs included above			
		£0				
	£0	£0	£57,097	£111,199	£48,444	£80,300
	0%	0%	19%	37%	16%	27%
DP4lib			£195,747	£90,103	£90,103	£151,192

Total Estimated Lifecycle Cost: **£297,040**

NB - not all DP4lib data entered

All figures on this sheet are rounded to 2 significant figures.

DP4lib £527,164

Overall, 56% of costs from the DP4lib model were able to be allocated within the LIFE3 model, leaving 44% unallocated. In comparing the processes undertaken in the two models, it was evident that within the DP4lib model the whole ingest process was run again. Given that this was the case, the ingest costs given above were doubled an assumption being that all sub-processes were the same for the re-ingest process. This gave a revised estimated lifecycle cost of € 417,882: Ingest - € 134,749 (32%), Bit-stream preservation - € 131, 215 (31%), Content Preservation - € 57,164 (14%), Access - € 94,754 (23%), as shown in Figure 10 below.

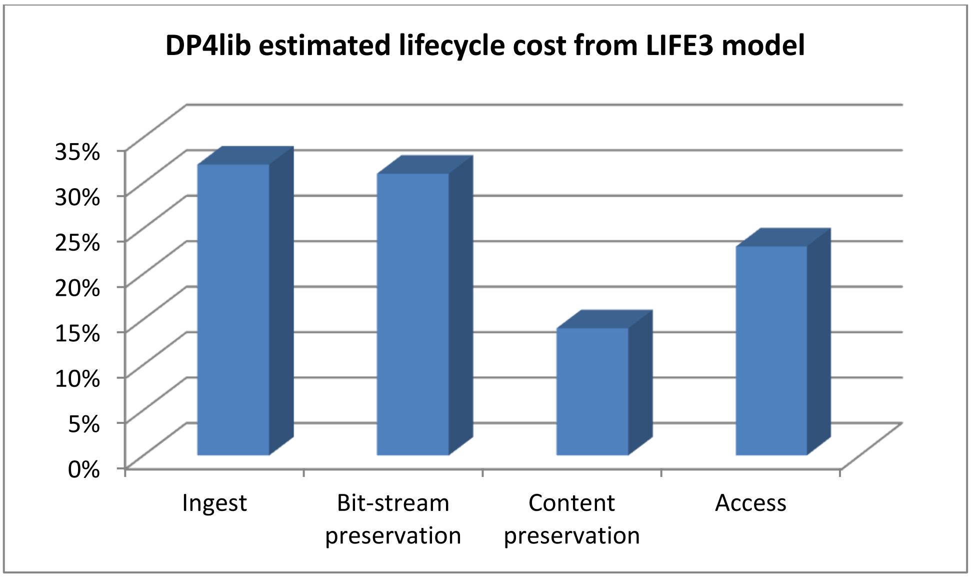


Figure 10: DP4lib test data in the LIFE3 model

Therefore, the final cost allocation was 67% which left 34% of costs unallocated from the DP4lib model.

The reason that costs could not be allocated was that there was no clear match between the processes or activities across the DP4lib and LIFE3 cost models. On analysing access costs, the Refine Access Data spread sheet allows for entry of staff costs only by document type with a unit of measure of, days of effort for staff at a particular grade. This unit is quite difficult to measure. A recommendation may be to simplify this stage and to allow for the entry of other costs as well as staff e.g. hardware and external services (as allocated from the DP4lib model). The following sub-processes could not be assigned:

- Auditing roles (customers can deliver in different roles) – this is statutory for German organisations and must be differentiated between the service for third parties and the effort for fulfilling the legal requirements. Here no mapping between this DP4lib activity and other activities is possible as the process is quite different
- Integrity checks are also different due to the fact that authors, publishers, service providers and external data centres are involved
- Classification of digital object - an evaluation of the probable long-term suitability of digital objects, through risk management which is undertaken

In the other cases either no mapping is found or other distribution keys are used, for example: unpack transfer packages, SIP handling, reporting (for customers), status report and contract negotiations.

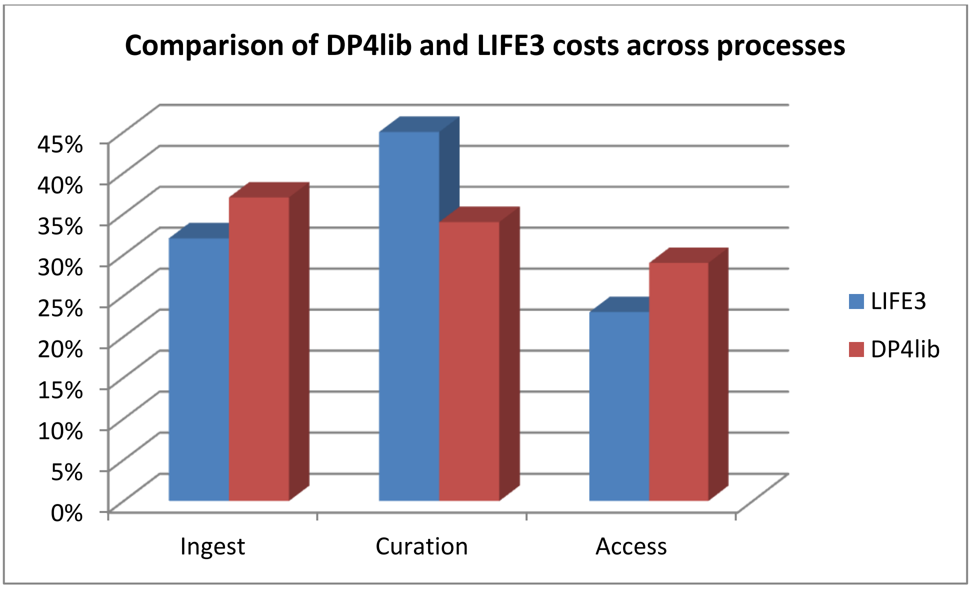


Figure 11: DP4lib and LIFE3 costs compared across processes

When comparing the allocation of costs (although only 67% of DP4lib costs) it is interesting to note the comparison across the three processes of ingest, curation and access (where bit-stream preservation and content preservation for LIFE3 have been included under curation processes). As shown in Figure 11 above DP4lib costs are higher across ingest and access processes whereas LIFE3 costs are higher under curation. It should be noted that this is not strictly a fair comparison as not all DP4lib costs are included.

3.4.4 Conclusion

An iterative approach was adopted. On initial basic input of costs this provided an allocation of 42% of costs from the DP4lib model to the LIFE3 model. Once detailed costs were entered allocation rose to 51% of total costs with the high level allocation providing a total of 56% of costs. When evaluated further, as the ingest process is carried out twice in the DP4lib model; there was a final allocation of 67% of costs between the two models. One third of costs could not be allocated due to processes not being the same across the organisations for which the models were created.

The LIFE3 model could be simplified at some data entry points especially where the unit of measure is too complicated to calculate, for example, days of effort for a staff member of a particular grade. Also for some elements of the LIFE3 model costs other than those related to staff (e.g. hardware and external services) should be allowed to be entered. These categories of cost could be easily added to the refine data work sheets. Additional case study data would improve the model across various file formats. In the case of data already included, average file sizes in terms of storage capacity could be reviewed, for example, web site storage sizes, these can be reviewed in light of recent statistics which are available from the web-archiving team at the British Library.

3.5 CONCLUSIONS OF COST MODEL TESTING

Overall, testing models through high level comparisons seems possible as shown by the three tests completed and provides evidence that cost models are quite specific to their own organisational requirements for the provision of activities, services or workflows. When underlying assumptions of the model are understood the exercise proves to be valuable in understanding the cost allocations and processes identified within costs models, even though they are quite specific to the organisations for which they were created. However, as costs modelled are so specific, differences occur in relation to preservation activities, breakdown of costs and apportionment and allocation methods (e.g. direct and indirect cost calculations). It should be noted that considerable effort is required to amend your own cost data so that it best fits the model being used or tested. This exercise can be time-consuming and may prove to be difficult as either the organisational data is incomplete or it is not possible to allocate costs as required by the model.

For any future testing or validation of the DANS and LIFE3 models it is advisable to use data from within the organisation for which the model was created. Where this is not possible, a good understanding of the model being used or tested as well as your own cost data is essential, for example, in relation to organisational structures and cost allocations, as no standard approach is currently available. As the DANS and DP4lib models are under development and also not yet available, it may be that when they do become available in the future they will provide transparency around the services being offered. Other repositories or archives may wish to offer similar services. Due to funding constraints, these types of organisations are looking at new sources of income, for example, the KB are considering this option. In these instances the DANS and DP4lib cost models would provide very useful tools to use when repositories or archives are looking at the costs of their own services as additional revenue streams. It is understood that the DP4lib model will provide a generic model which can be used digital repositories in libraries, archives or other heritage institutes and further validation and testing results for this model will provide clear evidence as to the generic nature of the model and demand for such a model exists.

Where new models are to be created by organisations, the generic features of each of the models tested might be adaptable. For example, the activity breakdown used by DP4lib might be preferred; or it might be more useful to use the direct/indirect costs distinction adopted by DANS; both DANS and DP4lib might be better suited strictly for services, whereas LIFE3 might be more widely applicable to digital preservation workflows.

4 MODELLING THE BENEFITS OF DIGITAL PRESERVATION IN RELATION TO COSTS

4.1 INTRODUCTION

ISO 16363 (para 3.4.3) requires that 'The repository shall have an on-going commitment to analyse and report on financial risk, benefit, investment, and expenditure'. This can be demonstrated in respect of benefits by including cost/benefit analyses in service planning and management documentation. The objective is to be able to show that the repository is 'maintaining an appropriate balance between risk and benefits, investment and return'.

Of the cost models considered here only KRDS and DANS incorporate any attempt to represent benefits within the model in relation to costs:

- KRDS2 describes a high-level Benefits Framework in its final report [5], as well as Benefits Framework [6] and Value Chain and Benefits Impact tools [7]
- The DANS activity-based costing model provides a Balanced Scorecard methodology that can be used to describe and report on benefits through its Success Factor and Performance Indicator categories [8].

A cost model describes only the relationship of activities to costs, and does not cover all aspects of the full business case, such as scope, constraints, assumptions, benefits and measures of success, quality management, options appraisal, value for money and risks. During the development of cost models it is presumed that the preservation rationale and benefits will have been separately developed, and will be implicit in the activities of the cost model.

But how in a given business case is the relationship between costs and benefits conceptualised and qualified, and how is the relationship defined within the overall business case? Due to the stage of development of cost models and their lack of availability and use this leaves open the question of how the costs of preservation activities can be related to their proposed benefits within the full business case. As Ross (2012) observes: 'benefits to be derived from long-term preservation have proved elusive and arguments which might convince commercially-minded business leaders of the benefits are restricted' [9].

4.2 OVERVIEW OF BENEFITS OF DATA PRESERVATION SUSTAINABILITY

The literature on the sustainability of data preservation provides little in the way of specification of benefits or correlation of benefits to the costs of preservation activities: most often, benefits are taken as assumed, and described, if at all, in generic terms with little practical relevance to any given case. The situation is well stated by Beagrie et al [5] .

Analysis of the costs of preserving research data sets is not enough to assess the economic feasibility of a particular digital preservation activity. Cost analysis should be accompanied by a framing of the benefits from preservation – in other words, the value that is anticipated to emerge from the investment in maintaining the long-run existence and accessibility of research data. Much of the literature addressing economic issues related to digital preservation focuses on the cost side of the cost/benefit equation. Comparatively little attention is paid to articulating the benefits to stakeholders arising from the preservation activity. Instead, the benefits conferred from investment in digital preservation often are either assumed to be common knowledge, or are expressed in terms far too generic to be of practical use for decision-making purposes (e.g., 'preserving society's digital record for future generations', etc.).

Brian Lavoie, co-Chair of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access (BRTF), has also expressed the need to analyse the nature of the cost-benefit relationship implicit in any digital preservation initiative, and to build economically-sustainable business models on this basis: 'Developing economic strategies that ensure a sustained flow of resources from those who require or benefit from digital preservation services, to those who are willing to perform digital preservation services, is an essential aspect of economical sustainability' [10].

The majority of the actors in digital preservation are what the BRTF refers to as 'proxy institutions': like museums and libraries and other cultural institutions, it is central to their mission to preserve in the interests of a general social benefit, which by its very nature cannot be expressed in terms of an economic rationale. Institutions such as the British Library and the National Archives, and other European libraries, are mandated by their governments to preserve for the future, and in so far as they exist to preserve, the case for preservation as such does not have to be made.

But in practice any preservation activity will be premised on a distinction between what is to be preserved, and what is not to be preserved, and will involve assumptions about what benefits are expected to accrue, who is expected to benefit, and the likely scale and extent of the benefit. These assumptions will inform operational cost-benefit assessments, along such lines as: is it worth keeping these data at this cost, given the likelihood of re-use and the cost of recreation; or, are these data sufficiently valuable to justify a given level of expenditure?

If these assumptions are not explicitly stated in such a way that the activities which follow from them can be undertaken consistently, accountably, and with due consideration for the interests of the beneficiaries of preservation, then the preservation service is not well managed and may be at risk of failure.

The process of working out the benefits to be gained from a preservation service enables an organisation to clarify its strategic objectives, define the scope of required activities, understand how preservation outcomes relate to costs and activities, and creates confidence on the part of organisational stakeholders in the value proposition represented by the preservation service. Keeping the benefits profile of a service under on-going review allows an organisation to manage costs and activities in relation to outputs, and to plan for and respond to changes in the service context, such as growth in the volume of data under management, evolution in the profile and needs of users, or a loss of funding.

The difficulty of modelling benefits is that they are not necessarily such as can be accommodated within an institutional business case: they are rarely realised within a neat five-year time horizon, and the beneficiaries are often not, or only minimally and indirectly, the institution undertaking the preservation activity or its researchers. Many benefits are indirect or diffuse and presume long-term investment. Nor can they be easily measured: at best one can employ proxies that may act as partial indicators of immediate benefit, for example: traffic to the data archive, numbers of registered users, dataset deposits and downloads, and revenue. Such measures cannot represent the future value of data, or the benefit of providing the data in terms of the research or economic value that would have been unrealised had the data had not been preserved in the first place.

4.3 BENEFITS ANALYSIS AND BUSINESS CASES: THREE EXAMPLES

The difficulties outlined above are reflected in the relative paucity of initiatives to conceptualise or develop models for digital preservation that formulate the relationships between preservation activities and costs and benefits. Below we outline three attempts to do this that merit consideration. The work of the Blue Ribbon Task Force on Sustainable Digital Preservation does not describe a preservation model, but takes a high-level approach to preservation strategy, which explains clearly the need to factor in the benefits side of the cost-benefits equation, and for this reason is valuable. Otherwise, KRDS and DANS provide the only well-realised preservation cost models that address their relationship to the broader business case for preservation and incorporate significant benefits components in defined relations to their cost components.

4.3.1 Blue Ribbon Task Force

The Blue Ribbon Task Force (BRTF) adopts a high-level strategic approach to the benefits of digital preservation and the incentives for organizations to undertake preservation activities [11]. Although wide-ranging in scope and largely addressing the policy framework for preserving digital information in its broadest sense, the Task Force does address the particular case of research data, and submits to analysis many of the often unexamined assumptions that underlie preservation in initiatives.

According to the BRTF, all digital preservation is characterised by four key ‘structural challenges’:

- Preservation takes place over long time periods, and does not neatly fit the business horizons of an organization;

- The beneficiary stakeholders are a diffuse and indeterminate set, and their function in the long-term sustainability of a preservation service cannot be computed;
- Preservation suffers from misaligned or weak incentives, because there is often a disjunction between 'communities that benefit from preservation (and therefore have an incentive to preserve), and those that are in a position to preserve (because they own or control the resource) but lack incentives to do so'. This is the so-called 'free-rider' problem, whereby 'the cost of preservation may be borne by one organization but the benefits accrue to many';
- There is often a lack of clarity about roles and responsibilities among stakeholders, causing systemic inefficiencies in the overall data preservation economy.

Because so much of the value proposition in preservation is potential and contingent in value, the Task Force suggests that benefits analysis can most usefully be concentrated on current user cases:

Preserving materials with clearly defined current uses implicitly creates the option of having the assets available for as-yet-unknown uses that may emerge in the future. Therefore, to the extent possible, value propositions should focus on the benefits generated for current users – about whom decision makers are understandably most concerned – rather than focusing too much on benefits to future generations and unknown future uses.

The benefits of preservation investments should always be framed in terms of the possible uses for digital assets; cases should seek to emphasize near-term uses, but look as far ahead as possible: these are principles any preservation service would do well to take into account in making a business case. But the Task Force does provide a reference to emerging outputs from the Economic Sustainability Reference Model under development, although this is not a cost model it looks at value in terms of economic sustainability.

There had been a proposal in the early days of the Task Force to develop a 'menu' of economic models for sustainable digital preservation based on a range of digital preservation scenarios: 'recommendations in the report might take the form "if your digital preservation context is similar to *X*, we recommend that you consider using economic model *Y* to organize your digital preservation activities in a sustainable way"' [10].This does not appear to have materialised in the final outputs of the Task Force, but something of this nature may emerge from the outputs of the Economic Sustainability Reference Model [12], an offshoot of the BRTF that is currently in development. Although this is not a cost model, it does conceptualise Value for stakeholders as a necessary condition for economic sustainability, and defines key entities on the Demand Side as Current Beneficiaries and Future Beneficiaries.

4.3.2 Keeping Research Data Safe

The Keeping Research Data Safe (KRDS) projects (which reported in 2008 and 2010) may be said to develop the idea of a practical benefits-oriented preservation planning approach outlined by the BRTF. KRDS is the preservation model that most systematically addresses preservation from the perspective of the benefits to be achieved. As the final report put it, 'when stating the direct benefits of digital preservation, the focus should be on the outcome from preservation, not the process' [5]. This outcomes-oriented approach is quite distinct from the approach adopted by most of the other cost models under consideration in this report, which are almost exclusively process-based.

This perspective is outlined in the first KRDS report [13], which derived four categories of benefits based on project interviews and discussions:

- Protecting investment in research: cost of preserving as against cost of irretrievable data loss or having to recreate data;
- Preserving opportunities for further research: researchers can validate and build on existing data;
- Promoting the work of the institution and the researcher: maximizing research impact and cultivating reputation;
- Supporting research and learning workflows as part of the 'general information infrastructure needed to support research and learning workflows at HEIs'.

The KRDS/I2S2 Digital Preservation Benefit Analysis [14] toolkit included the KRDS Benefits Framework Tool and Value Chain and Benefits Impact Tools which are formatted as spread sheet templates. The Benefits Framework is conceived as an 'entry-level' tool suitable for high-level strategic planning. It is meant to help preservation planners articulate the aims of their preservation activities in ways that can be easily communicated to stakeholders. According to the *Guide*, 'The Framework can assist in prioritizing alternative curation investments, and justifying data curation costs within funding applications'. The Value Chain and Benefits Impact Analysis Tool is more advanced, designed primarily for use in longer-term and intensive activities such as evaluation and strategic planning.

The Benefits Framework describes the benefits of preservation of research data in three dimensions: the outcome achieved; when the outcome is achieved; and who benefits from the outcome. Each dimension is further divided into two categories:

Dimension 1: What are the outcomes?

- Direct benefits: positive impacts obtained from investing in a data curation activity, such as new research opportunities; compliance with funders' data management mandates; providing evidential validation of research results;
- Indirect benefits: negative impacts avoided by investing in a data curation activity, such as avoiding the cost of re-creating data at a later time; reducing long-term preservation costs through up-front investment; protecting return on investment in data creation.

Dimension 2: When are the benefits received?

- Near-term benefits: benefits expected to be received up to five years from the present;
- Long-term benefits: benefits expected to be received beyond five years from the present.

Dimension 3: Who benefits?

- Internal benefits: benefits which impact stakeholders internal to/affiliated with the organisation undertaking the data curation activity;
- External benefits: benefits which impact stakeholders external to/not affiliated with the organisation undertaking the data curation activity.

Once these benefits have been identified and articulated within the framework, the organisation has a basis on which to identify measures of benefits achieved or to assign values to benefits according to organisational priorities, and to relate benefits to preservation activities and costs.

The Value Chain and Benefit Impact Analysis Tool allows this analysis to be carried out in more detail by relating benefits to specific activities or phases of activity. Value chain analysis is a technique that has been widely used in industry to analyse processes into component value-adding activities. This makes it easier to identify metrics and indicators for the value added by a given activity. The tool provides a schema for mapping benefits to activities, stakeholders, time to realisation, and impact measures (which may be qualitative or quantitative). The activities can be directly mapped onto the KRDS Activity Model, the project's costing model, so permitting direct mapping of benefits to activities and costs.

The *Guide to the Value Chain and Benefit Impact Analysis Tool* proposes a number of value and efficiency metrics (e.g. numbers of dataset publications/citations, cost/time savings), which can be applied as measures of benefit according to different stakeholder groups (institutions, researchers/research teams, research support services) or categories of use (scholarly communication and access).

The KRDS project is uniquely valuable in providing such an elaborated attempt to develop a set of analytic concepts on both the costs and the benefits sides of the equation. It situates both costs and benefits within a high-level strategic framework. In addition, the tools are exemplified by worked examples from a number of different disciplinary fields, which demonstrate how they can be applied to a particular case.

4.3.3 The DANS CMDA Balanced Scorecard

The Cost Model for Digital Archiving developed by DANS is the only other cost model here considered that builds in a benefits analysis, in this case using a well-established modelling tool, the Balanced Scorecard (BSC) [15]. The BSC 'translates an organisation's mission and existing business strategy into a limited number of strategic business objectives that can be linked and measured operationally'.

As the Balanced Scorecard is used in DANS, the strategic objectives of the organisation are described as Success Factors, in turn qualified by a set of Performance Indicators. For example, the key long-term interests of DANS are expressed in two Success Factors: 'sources of revenue are sustainable', and 'there is a growth in the number of its supporters'. These are clearly measurable, and promote the general benefit DANS seeks under its mission statement Success Factor: 'to improve the research and data infrastructure in the social sciences and humanities'.

Success Factors and Performance Indicators can be mapped onto the DANS Activity-based Reference Model, allowing measured benefits to be related to activities and costs and providing for system feedback: if an expected benefit is not being realised at the required level, the model allows managers to track back to activity and cost inputs and make adjustments. In similar fashion, impacts on benefits of changes in activity or cost levels can be modelled.

As this summary makes clear, the DANS model is more focused on business process, while KRDS is addressed to a broader strategic framework. But incorporation of the Balanced Scorecard into the DANS model provides a model for organisations to translate broader strategic aims into specific business objectives: for example, the DANS mission Success Factor of being a leader in growth and innovation is expressed in terms of activities with given outcomes and defined costs.

4.4 SUMMARY: RELATING COSTS AND BENEFITS IN THE BUSINESS CASE

While there are a number of well-developed lifecycle and activity cost models available for the preservation of digital research data, as demonstrated by this report, by and large these have been presented in isolation from the modelling of benefits and mapping them to activities and costs within the full business case. Of the cost models considered in this report, only KRDS and DANS provide the benefits of preservation modelled and mapped to costs and activities. These represent a good foundation on which to build, but also highlight the fact that a cost model is of little use in isolation from a full business case when addressing the strategic objectives of the organisation.

This matters, because, in the words of Grindley (2013), organisations 'don't yet have the means or experience to accurately gauge the value that digital assets represent to their organisation, and consequently don't know to what extent digital preservation will serve the long-term financial and strategic interests of the organisation. The question thus becomes, "what is the business case for digital preservation?" [16]. Grindley believes that many digital preservation services are ill-equipped to make the case for investment to their managers and funders, and to establish criteria of success against which their service can be measured. Cost models by themselves will not make the argument; they can only be effective within a full business case.

Many of those organisations undertaking data preservation do so in accordance with an organisational mission to preserve cultural memory, support the scholarly community, or generate social or economic impact, and may not need to make the case for preservation as such. Even so, those who manage and fund data preservation services, if they are to do so effectively, have to understand the benefits they create for their stakeholders, the value of those benefits, and how they relate to service costs – and must be able to communicate this information to their stakeholders.

Any preservation activity will be informed by decisions about the scope of material to be preserved, the scale and quality of preservation activities, and how preserved assets will be made available for use. All of these decisions have cost implications, and require an understanding of the underlying business case. Cost and benefit contexts are complex and dynamic: funding levels and priorities, service costs, and stakeholder requirements all change over time. Service funders and managers who are informed about the preservation service and its component activities are enabled to manage and develop the service in a way that is cost-effective while maximising stakeholder value.

The business case will in essence provide two key sets of information: about the investment required (the cost), and value to be realised (the benefits). A well-formed business case will articulate the relationships between these two sides of the equation in terms that are rational and convincing. Where benefits are concrete, measurable in the same terms as costs, and accrue directly to the cost-bearer within defined time periods, as in the case of financial return, expressing the relationships between costs and benefits is relatively straightforward. But such neat equivalences are by and large not available to those institutions that undertake digital preservation. Benefits are often indeterminate, contingent and remote; even when they or their proxies can be measured, they are expressed in terms incommensurable with those used to express costs; furthermore, they may accrue largely to stakeholders other than the cost-bearer, and will accumulate over an indefinite duration.

These are challenges, but they do not preclude the development of robust analytic business models for data preservation services. In fact the espida project [17], a JISC-funded project (2007) to develop a strategic model for the expression of the costs and benefits of the preservation and disposal of digital assets in higher education institutions, has directly addressed the challenges faced by publicly-funded preservation services in developing a tool for project proposers and decision makers in higher education to build and analyse business cases for the preservation of digital assets [18]. Acknowledging that a large body of work on the costs of digital preservation exists, the espida project team place particular emphasis on articulating the benefits of preservation activities within the context of an organisation’s strategic objectives and business processes.

The espida project includes in its suite of tools outcome scorecards and cost templates, thus providing the means for an organisation to express the relationships between costs and benefits in terms of the business case. Although the espida project case studies are of institutional repositories and a museum archival collection, DANS in effect uses the espida approach to build a fully-fledged data preservation service business model that could be adapted and built on by other services.

4.5 CONCLUSION

There is undoubtedly a demand for full business models for data preservation that incorporate cost models, whether those considered in this report or others. The espida project affords a useful example of a business case framework that has been tried and tested in the context of a data preservation service, but there is room for other models to be developed and tested. There are well-established examples of business case best practice for public sector institutions [19], and many organisations will have their own business case guidelines which could be applied to data preservation services. As is the case with cost models, different kinds of business case will suit different organisations and service contexts: for this reason there would be value in having more published business models for digital preservation with practical tools and detailed case studies.

5 RELATIONSHIP WITH CO-ORDINATION ACTION, 4C PROJECT

The results of the work carried out in this work package and presented in this deliverable, may provide valuable results for the 4C project [20]. The findings of the mapping exercise provide areas for further investigation and development of cost parameters. The clarification of parameter definitions may be a valuable exercise to carry out and is an area which the 4C project will be looking at.

The results provided from testing the DANS, DP4lib and LIFE3 cost models provide a greater understanding of the way in which each of these cost models has been created. In some instances recommendations are provided to assist with any future development work. For example, the DANS model reflects the organisation which is very dynamic and has gone through organisational changes. The issues around how the cost model can be adapted to various needs and the barriers to using cost models are to be reviewed within the 4C project.

Early indications of the impact of our report are that the mapping exercise to ISO16363 is an interesting and useful methodology. This approach could result in the way that constraints and drivers that act upon the cost concept model are formulated within the 4C project.

A formal handover to the 4C project, of the results of the research undertaken within the APARSEN project, will take place within a joint workshop at the iPRES conference 2013 in Lisbon, Portugal.

6 THE FUTURE OF COST MODELS

As mentioned the 4C project will continue work in this field and provide an online curation costs exchange tool. For cost models to be applicable across a number of digital repositories and archives is inextricably linked to organisational cost structures and digital preservation workflows. Recommendations for future work are include within the report and there is certainly a need for further work around advocacy, advice and guidance around cost modelling. It may be that in terms of technical applications, simulation models could be reviewed, benchmarking exercise which allow for costs to be to be identified and compared for specific activities may be useful. There a number of areas of further development which need to be addressed and the need for developing cost models is evident – producing such a model which can be used in a standardised way is the next challenge.

7 REFERENCES

[1] D32.1 Report on cost parameters, http://www.alliancepermanentaccess.org/index.php/aparsen/

[2] Audit and certification of trustworthy digital repositories, ISO 16363 http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=56510

[3] OAIS reference model, ISO14721 http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=57284

[4] Mantid – software developed by Tessella which provides a similar function to a migration tool, in that it can be used to convert old RAW data into NEXUS format. Also allows sharing of data views and graphs

[5] Beagrie, N., Lavoie, B., and Woollard, M. (2010), Keeping Research Data Safe 2. Final Report. http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf

[6] Beagrie, N. (2011a), *A Guide to the KRDS Benefits Framework.* http://www.beagrie.com/KRDS_BenefitsFramework_Guidev3_July%202011.pdf

[7] Beagrie, N. (2011b), *A Guide to the KRDS Value-Chain and Benefit Impact Analysis Tool.* http://www.beagrie.com/KRDS_ValueChainImpactTool_Guide_v1July2011.pdf

[8] Palaiologk, A. S., Economides, A. A., Tjalsma, H. D. and Sesink, L. B. (2012), 'An activity-based costing model for long term preservation and dissemination of digital research data: the case of DANS,' International Journal on Digital Libraries, 12(4), 195-214. doi:10.1007/s00799-012-0092-1

[9] Ross, S. (2012), 'Digital preservation, archival science and methodological foundations for digital libraries', in *New Review of Information Networking*, 17(1), 43-68. http://dx.doi.org/10.1080/13614576.2012.679446

[10] Lavoie, B. F. (2008), 'The fifth blackbird: some thoughts on economically sustainable digital preservation', in *D-Lib magazine*, 14(3/4). http://www.dlib.org/dlib/march08/lavoie/03lavoie.html [1]

[11] Berman, F., Lavoie, B., Ayris, P. et al. (2010), Sustainable Economics for a Digital Plant: Ensuring Long-Term Access to Digital Information. Final Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access. http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

[12] Economic Sustainability Reference Model blog by Chris Rusbridge http://unsustainableideas.wordpress.com/economic-sustainability-ref-model-page/

[13] Beagrie, N., Chruszcz, J., and Lavoie, B. (2008), *Keeping Research Data Safe: A Cost Model and Guidance for UK Universities*. *Final Report.* http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf

[14] KRDS/I2S2 Digital Preservation Benefit Analysis Tools Project. http://beagrie.com/krds-i2s2.php. Accessed 11 June 2013.

[15] Kaplan, R. S. and Norton, D. P. (1992), 'The balanced scorecard – measures that drive performance', in Harvard Business Review, 70, 58-63. http://hbr.org/1992/01/the-balanced-scorecard-measures-that-drive-performance/ar/1

[16] Grindley's point is made in reference to funding recently provided by JISC in support of projects seeking to build the evidence base for on-going investment in digital preservation

[17] Espida Project. http://www.gla.ac.uk/services/library/espida/index.shtml. Accessed 11 June 2013

[18] Currall, J. and McKinney, P. (2007), espida Handbook. Expressing Project Costs and Benefits in a Systematic Way for Investment in Information and IT. http://hdl.handle.net/1905/691

[19] For example: The UK Government's Five Case Model, widely used in the UK public sector:

https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/190601/Green_Book_guidance_public_sector_business_cases_using_the_Five_Case_Model_guidance.pdf

[20] A Collaboration to Clarify the Costs of Curation, 4C project: http://4cproject.net/

[21] Open Planets Foundation, website, http://www.openplanetsfoundation.org/

8 BIBLIOGRAPHY

Beagrie, N. and Pink, C. (2013). *Research Data Management* and REF2014. http://opus.bath.ac.uk/35518/

Currall, J. and McKinney, P. (2006), *'Investing in value: a perspective on digital preservation'*, in D-Lib magazine, 12(4). http://www.dlib.org/dlib/april06/mckinney/04mckinney.html

Grindley, N. (2013), 'Building the business case for digital preservation', in Duranti, L. and Shaffer, E. (eds), *Proceedings of The Memory of the World in the Digital Age: Digitization and Preservation. An International Conference on Permanent Access to Digital Documentary Heritage*. 26-28 September 2012, Vancouver, British Columbia, Canada, 269-277. http://www.unesco.org/new/en/communication-and-information/events/calendar-of-events/events-websites/the-memory-of-the-world-in-the-digital-age-digitization-and-preservation/

Kejser, U. B., Nielsen, A. B. and Thirifays, A. (2013), 'Modelling the costs of preserving digital assets', in Duranti, L. and Shaffer, E. (eds), *Proceedings of The Memory of the World in the Digital Age: Digitization and Preservation. An International Conference on Permanent Access to Digital Documentary Heritage*. 26-28 September 2012, Vancouver, British Columbia, Canada, 529-539.

Rosenthal, D. S. H., Rosenthal, D. C., Miller, E. L. et al. (2013), 'The economics of long-term digital storage', in Duranti, L. and Shaffer, E. (eds), *Proceedings of The Memory of the World in the Digital Age: Digitization and Preservation. An International Conference on Permanent Access to Digital Documentary Heritage.* 26-28 September 2012, Vancouver, British Columbia, Canada, 513-528.

ANNEX 1: COST MODELS – FURTHER DETAILS

As presented in D32.1 and included here for completeness.
Reference: Open Planets Foundation [21] pages, as updated by Paul Wheatley and Andy Jackson
CET – Cost estimation toolkit

- Estimates life cycle costs for scientific data activities, can potentially be applied to long-term archive systems
- Two excel based tools developed, CET software package is available http://opensource.gsfc.nasa.gov/projects/CET/index.php
- Paper published http://www.pv2007.dlr.de/Papers/Fontaine_CostModelObservations.pdf
- Developed by NASA and SGT
- Further reading:

  Fontaine, K., Hunolt, G., Booth, A., and Banks, M., 2007, Observations on Cost Modelling and Performance Measurement of Long Term Archives in PV2007 Conference Proceedings: http://www.pv2007.dlr.de/Papers/Fontaine_CostModelObservations.pdf

CMDP - Cost Model for Digital Preservation

- Estimates the costs of digital preservation (ingest, preservation planning and migrations, and archival storage), covers cultural heritage organisations
- Still under development, tool available
- Available on-line http://www.costmodelfordigitalpreservation.dk/
- Developed by the Royal Library of Denmark and the Danish National Archives
- Further reading:

  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). doi:10.2218/ijdc.v6i1.186: http://www.ijdc.net/index.php/ijdc/article/view/177

  Kejser, U., Nielsen, A., & Thirifays, A. (2009). The cost of digital preservation. Committee for Digital Preservation, Project Report v.1.0. Danish Ministry of Culture: http://www.costmodelfordigitalpreservation.dk/contact/cmdp-1---preservation-planning-and-digital-migrations/documentation

  Kejser, U., Nielsen, A., & Thirifays, A. (2011). Costs of digital preservation. Project Report for Phase 2: http://www.costmodelfordigitalpreservation.dk/contact/cmdp-2---ingest-and-archival-storage/documentation

DANS cost model

- Calculates the costs of archiving datasets, based on activity based costing and balanced scorecard, covers research data archives
- Validation to be undertaken
- Paper published on the model http:/www.springerlink.com/content/v3r1282x328m607m//?MUD=MP
- Developed by DANS, Data Archiving and Network Services, Netherlands
- Further reading:

  Palaiologk, A. S., Economides, A. A., Tjalsma, H. D., & Sesink, L. B. (2012). An activity-based costing model for long-term preservation and dissemination of digital research data: the case of DANS. International Journal on Digital Libraries, 12(4), 195–214. doi:10.1007/s00799-012-0092-1: http://link.springer.com/article/10.1007%2Fs00799-012-0092-1

DP4lib - Digital Preservation for libraries

- Calculates costs by a service model for long term preservation services to third parties, covers any sector
- Validation taking place this year, documentation and a calculation spread sheet are available on the following website: http://dp4lib.langzeitarchivierung.de/index_downloads.php.de.
- Developed by the DNB
- Further reading:
  Paper published on the model
  http://aparsen.digitalpreservation.eu/pub/Main/CostModels/DP4lib-Cost-By-Service-CostModel.docx

ENSURE project

- Estimates costs of digital preservation activities, assumes cloud storage is used, covers healthcare, clinical trials and financial sector, may be extended to manufacturing sector
- Initial model to be developed further
- Being developed by EC FP7 project, ENSURE (Feb '11 – Jan '14) http://ensure-fp7-plone.fe.up.pt/site
- Further reading:
  M Badawy, E Shehab, P Baguley, MD Wilson "Towards a cost model for long term digital preservation", In Proc. 2012 ISPA/SCEA Joint International Conference: Assuring cost efficiency : global solution, Brussels, Belgium, 14-16 May 2012: http://epubs.stfc.ac.uk/bitstream/7711/Towards%20a%20Cost%20Model%20for%20Long%20Term%20Digital%20Preservation.pdf

ISIS facility model

- Applied specifically to long term preservation costs of data from ISIS facility at STFC (scientific research data)
- Not applicable to other areas
- Developed as part of Cranfield University MSc project in collaboration with STFC
- Further reading:
  Poster published http://ensure-fp7-plone.fe.up.pt/site/Poster.pdf

KRDS – Keeping research data safe (KRDS + KRDS 2)

- As well as activity cost model provides lists of benefits and potential metrics for research data, is applicable more widely.
- Toolkits - benefits analysis, value and impact - for proposals, evaluation and planning
- Published factsheet, user guide http://www.beagrie.com/krds.php
- Development of toolkits funded by JISC partners in project include Charles Beagrie Ltd, UKOLN, DCC, UCL, UKDA, ADS, OCLC
- Further reading:
  Beagrie, N., Chruszcz, J., & Lavoie, B. 2008. Keeping Research Data Safe. A Cost Model and Guidance for UK Universities, Copyright HEFCE 2008: http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf70H
  Beagrie, N., Lavoie, B., and Woollard, M. (2010), Keeping Research Data Safe 2: http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf

LIFE3 – Life Cycle Information for E-literature

- Looks at long-term costs of digital preservation for DP repositories
- Third phase of the LIFE Project producing a predictive costing tool (not developed fully), excel version is available for use
- Published excel tool and papers http://www.life.ac.uk/
- Developed by UCL and BL, project funded by JISC and RIN
- Further reading:

  Wheatley and Hole. LIFE3: Predicting Long Term Digital Preservation Costs. In iPRES2009, September 2009: http://www.fsl.cs.sunysb.edu/docs/unesco12/UNESCO2012-storage-econ.pdf

  Hole, Brian, Li Lin, Patrick McCann, and Paul Wheatley (2010), "LIFE3: A predictive costing tool for digital collections," 7th International Conference on Preservation of Digital Objects (iPRES2010), Vienna, September 19-24, 2010: http://www.life.ac.uk/3/docs/Ipres2010_life3_submitted.pdf

PrestoPRIME – cost model for digital storage

- Provides cost information and long term forecasting for mass digitisation of AV materials
- Tools available and still under development
- Published report http://PrestoPRIME.it-innovation.soton.ac.uk/planning-tool/accounts/login?next=/planning-tool/
- Developed within EC FP7 project http://www.PrestoPRIME.eu/
- Further reading:

  Addis, Matthew and Jacyno, Mariusz (2010) *Tools for modelling and simulating migration based preservation.* PrestoPRIME Consortium, 56pp: https://prestoprimews.ina.fr/public/deliverables/PP_WP2_D2.1.2_PreservationModellingTools_R0_v1.00.pdf

  Westerhof, Hans, Ubois, Jeff and Snyders, Marius (2011). Financial Models and Calculation Mechanisms. PrestoPRIME Consortium, 35pp: https://prestoprimews.ina.fr/public/deliverables/PP_WP6_D6.3.1_FM_calculation_R0_v1.01.pdf

May be of interest:

OECD – International Standard Cost Model Manual

- Determines administrative costs, provides transparent measures
- Developed by the Standard Cost Model Network
- Published manual http://www.oecd.org/regreform/regulatorypolicy/34227698.pdf

ANNEX 2: DETAILED MAPPING ANALYSIS

Organisational Infrastructure

Cost Model	Governance, organisational viability			Organisational structure and staffing	Procedural accountability, preservation policy						Financial sustainability			Contracts, licenses and liabilities	
Sub- heading	Mission Statement on the preservation of digital information	Preservation Strategic Plan	Collection Policy	Staff with adequate training and skills	Defined Designated community and knowledge base	Preservation Policy to ensure Preservation Strategic Plan met	Audit trail of changes to operations, procedures, software, hardware	Transparency and accountability in all actions supporting the operation and management of the repository that affect the preservation of digital content over time.	Information integrity measures to be defined, collected and tracked	Self assessment, external certification	Business planning processes	Financial practices and procedures, 3rd party audits	Analyze and report on financial risk, benefit, investment, and expenditure (including assets, licenses, and liabilities)	Contracts or deposit agreements for digital materials	Intellectual property rights mgt
CMDP				X			x								
DANS				X										X	
DP4lib				X				X	X		X	X	x	X	X
ENSURE															
ISIS															
KRDS		x	x	X	X					X	x	x	X	x	
LIFE3	X	X	X	X	X	X	x	X	X		x	x	x	X	X
PrestoPRIME		x			x				X		x		x		

Digital Object Management

Cost Model / Sub- heading	Ingest: acquisition of content								Ingest: creation of AIP										Preservation planning				AIP preservation		Information management			Access management	
	Content Information and the Information Properties	Content Information on deposit	Specifications enabling recognition and parsing of the SIPs.	Verification of Producer identity	Ingest process to verify SIP	Preservation control over the Digital Objects	Responses to producer/depositor	Content acquisition records	Definition for parsing AIP and for long term preservation	AIP construction form SIP	Document SIPs	Persistent unique identifiers for AIPs	Provide Representation Information (RI) for all digital objects	Preservation Description Information (PDI) for Content Information (CI)	AIP Content Information understandable to Designated Community	AIP verification	Repository integrity verification	AIP creation records	Preservation strategy	Monitor preservation environment	Preservation plan changes	Preservation activity effectiveness	AIP storage spec	AIP storage and preservation records	Discovery and identification of material by Designated Community	AIP descriptive info	Bi-directional linkage between AIP and descriptive info	Access policies	Authenticity of Digital Objects
CMDP	X	X		X	X	X	X			X				X		X	X	X	X	X	X		X	X				x	
DANS		X			X	X		X			X		X						X	X	X				X	X	X		
DP4lib	X	X	X	X	X	X	X	X	X	X	X	X	X	X			X	X					X	X	X	X	X	X	X
ENSURE		X	X	X	X			X	X	X	X	X	X	X	X	X			X	X	X	X	X	X	X	X	X	X	
ISIS		X											X		X					X				X		X			
KRDS	X			X	X	X	X	X												X					X			X	X
LIFE3	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			X		X	X	X	X	X	X	X
Presto PRIME					x																				x			X	

Infrastructure and security risk management							
Cost Model	Subheading	Technical infrastructure risk management		Security risk management			
		Identify and manage risk management to preservation operations and system infrastructure	Manage number and location of copies of Digital Objects	Analysis of security risk factors associated with data, systems, personnel, and physical plant	Security risk controls	Staff roles and responsibilities	Disaster recovery plans, one off-site backup and offsite recovery plan
CMDP		X	x	X	X		X
DANS					X	X	X
DP4lib			X				
ENSURE		X	X				
ISIS		X					
KRDS		X	X				
LIFE3		X	X				X
PrestoPRIME		X		X			x

Key: x = cost parameter partially mapped
X = cost parameter fully mapped