Collaboration to Clarify the Cost of Curation

D3.1—Evaluation of Cost Models and Needs & Gaps Analysis

Deliverable Lead:	Det Kongelige Bibliotek (KBDK)
Related Work package:	WP3—Assessment
Author(s):	Ulla Bøgvad Kejser (KBDK) Kathrine Hougaard Edsen Johansen (DNA) Alex Thirifays (DNA) Anders Bo Nielsen (DNA) David Wang (SBA) Stephan Strodl (SBA) Tomasz Miksa (SBA) Joy Davidson (HATII-DCC) Patrick McCann (HATII-DCC) Jaan Krupp (NLE) Heiko Tjalsma (KNAW-DANS)
Dissemination level:	Public
Submission date:	30 June 2014
Project Acronym:	4C
Website:	http://4cproject.eu
Call:	FP7-ICT-2011-9
Project Number	600471
Instrument:	Coordination action (CA)—ERA-NET
Start date of Project:	01 Feb 2013
Duration:	24 months

Project funded by the European Commission within the Seventh Framework Programme		
Dissemination Level		
PU	Public	✓
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Version History

Version	Date	Changed pages / reason	Modified by
0.01	21st June 2013	First draft	KHEJ
0.02	23rd June 2013	Section on "Economic models in the field of digital curation" added	KHEJ
0.03	2nd October 2013	Gap analysis, draft, added	AT
0.04	17th October 2013	Draft report for QA	AT
0.05	21st October 2013	First pass edit and QA	PLSS
0.06	28th October 2013	Draft report MS 12 for QA (SBA revision of section 5.2.2 and section 6)	UBK
0.07	28th October 2013	Second pass edit and QA	PLSS
0.08	30th October 2013	Corrected section 6 and 7 for QA	UBK
0.09	31st October 2013	Corrected section 2 and 3 from ABN	UBK
0.10	5th November 2013	Integration of input from HATII-DCC (Chapter 5)	UBK
0.11	6th November 2013	Third pass edit and QA	PLSS
0.12	12th November 2013	Fourth pass edit and QA	PLSS
0.13	5th December 2013	Integration of partner comments and edit	PLSS
0.14	31th January 2014	Streamlining and integration of partner/ community comments	UBK
1.00	31st January1 2014	Released version	PLSS
1.01	10th March 2014	Logo Update	PLSS
1.02	10th May 2014	Post review revision	PLSS

Acknowledgements

This report has been developed within the project “Collaboration to Clarify the Cost of Curation” (4cproject.eu). The project is an ERA-NET co-funded by the 7th Framework Programme of the European Commission.

The 4C participants are:

Participant organisation name	Short Name	Country
Jisc	JISC	UK
Det Kongelige Bibliotek, Danmarks Nationalbibliotek og Københavns Universitetsbibliotek	KBDK	DK
Instituto de Engenharia de Sistemas e Computadores, Investigacao e Desenvolvimento em Lisboa	INESC-ID	PT
Statens Arkiver	DNA	DK
Deutsche Nationalbibliothek	DNB	DE
University of Glasgow	HATII-DCC	UK
University of Essex	UESSEX	UK
Keep Solutions LDA	KEEPS	PT
Digital Preservation Coalition Limited by Guarantee	DPC	UK
Verein Zur Forderung Der It-Sicherheit In Osterreich	SBA	AT
The University of Edinburgh	UEDIN-DCC	UK
Koninklijke Nederlandse Akademie van Wetenschappen -Knaw	KNAW-DANS	NL
Eesti Rahvusraamatukogu	NLE	EE

Authors:	Ulla Bøgvad Kejser (KBDK) Kathrine Hougaard Edsen Johansen (DNA) Alex Thirifays (DNA) Anders Bo Nielsen (DNA) David Wang (SBA) Stephan Strodl (SBA) Tomasz Miksa (SBA) Joy Davidson (HATII-DCC) Patrick McCann (HATII-DCC) Jaan Krupp (NLE) Heiko Tjalsma (KNAW-DANS)
Editor	Paul Stokes (Jisc)

Table of Contents

Figures

Tables

Executive Summary

This report *'D3.1—Evaluation of Cost Models and Needs & Gaps Analysis'* provides an analysis of existing research related to the economics of digital curation and cost & benefit modelling. It reports upon the investigation of how well current models and tools meet stakeholders' needs for calculating and comparing financial information. Based on this evaluation, it aims to point out gaps that need to be bridged in order to increase the uptake of cost & benefit modelling and good practices that will enable costing and comparison of the costs of alternative scenarios—which in turn provides a starting point for a more efficient use of resources for digital curation.

To facilitate and clarify the model evaluation the report first outlines a basic terminology and a general description of the characteristics of cost and benefit models.

The report then describes how the ten current and emerging cost and benefit models included in the evaluation were identified and provides a summary of each of the models. To facilitate comparison of the models, it also provides tables that lists each of the models' core features, such as which information assets they handle, which curation activities they address and how they breakdown costs.

This is followed by an in depth analysis of stakeholders' needs for financial information derived from the 4C project stakeholder consultation.

The stakeholders' needs analysis indicated that models should:

- support accounting, but more importantly they should enable budgeting
- be able to handle various types and amounts of assets and use cases
- have a sound definition and breakdown of costs and enable modelling of cost variables
- be supported by easy to use tools, preferable with default values/settings
- support assessment of the benefits and value of digital curation

These needs were used to inform an analysis of the previously identified models to identify gaps in the models capabilities to meet user needs (as defined by the stakeholder analysis). The most important gaps we discovered were associated with:

- poor usability
- lack of reliability
- lack of consensus on how to define, qualify and structure cost data
- lack of representation of the benefits of the investments in digital curation

On top of the needs and gap analysis, we identified actions which, if practical and possible may support the uptake and inform further development of cost & benefit models within the digital curation field. These recommendations for investigation and action include:

- provision of a high-level quick entry guide to all existing models that describes the scope and structure of the models indicating their relevance for different stakeholders and use cases
- provision of a vocabulary and a generic description of cost & benefit models
- provision of clearly designed and user-friendly tools with default reference settings that can be fine-tuned to accommodate for various stakeholder needs, and usable user-interfaces
- provision of benefit models in addition to the cost models
- provision of a shared knowledgebase with cost data and use cases

Finally, we have developed a set of good practices proposals for developers of models within the field of digital curation, and here, as in many other areas, the most important point is to keep it simple.

1 Introduction

Sustainability is a key issue for a wide range of private and public organisations responsible for managing digital information assets such as business records, research data, cultural heritage collections, personal archives and other assets that represent value to the organisations and others. To ensure timely funding, the organisations need to understand the economic lifecycle that they operate in and the costs and benefits that the assets incur or engender. Likewise, suppliers of asset management systems and services need to have detailed knowledge on what management activities are involved, how much they cost and what the cost drivers are. They also need to understand how the systems and services generate value for customers. This knowledge and understanding of costs and benefits supports the streamlining of businesses, increases in cost effectiveness and improves measurements of performance.

Stakeholders depend on the availability of sound financial information for accounting and budgeting to underpin this understanding. They must know the factual costs, for example records of the capital and labour costs required to develop and operate a specific system. But to understand the implications of the costs they must also have contextual information that describes the underlying assumptions about what is being priced, for example the specifications of the quality of a system (parameters such as how rigorous is the applied quality control, how well does it support ingest of different types of metadata and so on) and indications of the value that the system represents to different stakeholders. On the one hand this financial information allows financial transactions to be recorded and analysed for internal management purposes (and possibly for legal purposes as well). On the other hand it can also provide a basis for comparing solutions and thus support decision-making.

Costing digital asset management— digital asset management is also known as digital curation—is not a trivial task for a number of reasons, not the least of which is that there are many interrelated activities involved in curation. What's more, these activities can be implemented in many different ways and they can be set up to meet different quality requirements. This complexity makes it hard to specify the activities in a precise and clear-cut way. Also, cost models require detailed information for their calculations and often that information is intertwined with that of other cost centres. Indeed, there are no standardised ways of breaking down and accounting for the cost of curation activities. On top of this, digital curation activities depend heavily on constantly evolving technologies, which in turn leads to repeated changes in systems and procedures, and thus also in the costs. Assessment of benefits of digital curation is even less explored.

Assessing the costs and benefits of digital curation is not a new challenge per se, but coupled with the rapid growth in the amount and complexity of information assets, budgets for curation are increasingly under pressure and this has emphasised the need for reliable and comparable financial information to know where efficiencies can be achieved. This is where cost and benefit models come into play. Over the last decade several models have been developed to help organisations assess the costs and benefits of digital curation.

Awareness about the importance of costing the totality of digital curation—costing more than just storage costs—developed throughout the 1990s. This awareness was, amongst others, put into words in the report "*Preserving Digital Information*" [Garrett & Waters, 1996, p. 30]:

> *'In addition to managing their operating environment and the migration of information through hardware and software platforms, a third function by which digital archives fulfil their commitment to preserve electronic information is in managing the costs of these activities.'*

Around the turn of the millennium the first models for costing digital curation were developed. Through the 2000s to the present day a number of additional cost and benefit models for digital curation have emerged. An overview of models and bibliographies can be found at the Open Planets Foundation website[1], in a blog post from the Library of Congress "The Signal" blog[2] and in a deliverable report by the 4C project [4C, D2.1, 2013].

It is notable that institutions have seemed to find it easier to develop new cost benefit models as opposed to modifying and then reusing the existing ones. This has resulted in a relatively large number of different models. Although the models are tailored to specific needs there are some similarities. Of particular interest is the connection to the Reference Model for an Open Archival Information Systems (OAIS) [CCSDS, 2012 (ISO14721:2012)] that forms the basis of most existing models relevant to digital curation. In spite all the effort being put into research in the economics of digital curation there is still no consensus on the optimum way to model it.

Today's trends are towards developing a unified theory of how to model the costs and benefits of digital curation, and to make models more standardised [Lunghi et al., 2012, p. 195-268]. Alignment of methodologies will help facilitate comparison of alternative scenarios and selection of best practices to ultimately gain efficiencies in digital curation.

This project, the Collaboration to Clarify the Costs of Curation (4C), being a coordination action project, follows these trends and seeks to bring together and utilise existing knowledge in the field of the economics of digital curation, that being the specific goal of this report of deliverable D3.1 to "*Evaluate existing cost models and produce a needs and gap analysis*".

The next section explores this working title, a title and task which require some unpacking to clarify their scope and formulate their aims more explicitly. See also Appendix A.1—Terms and definitions for an explanation of the domain related nomenclature.

1.1 Scope

The first part of the stated purpose of the task is to "*Evaluate existing cost models*," and cost modelling in the field of digital curation is indeed the main focus of this work. However, it is also fully acknowledged by the digital curation community that costs are inextricably interwoven with the benefits and value that they bring, and therefore the scope of the task also includes evaluation of benefit models. More precisely, the evaluation includes an overall description of the purpose of existing models, as well as an in depth analysis of selected model characteristics, such as how the models structure cost, how they model variables, and how usable they are. Thus, the evaluation will describe the individual models strengths and weaknesses and seek to identify best practice—effective ways of modelling costs and benefits that more users can apply, and which can also be used as a benchmark for improving modelling methods further. Again, the evaluation goes into detail with cost models, and only deals with benefit models at the high level.

In the 4C deliverable 'D4.3—Quality and trustworthiness as economic determinants in digital curation' digital curation is defined as being *"…about ensuring that digital objects remain usable*"[3]. For the sake of clarity—and also to encourage a degree of standardisation in the field of digital curation—we have chosen

[1] Open Planets Foundation, Digital Preservation and Data Curation Costing and Cost Modelling, http://wiki.opf-labs.org/display/CDP/Home [Accessed 31 Jan 2014]
[2] The Signal – Digital Preservation, A Digital Asset Sustainability and Preservation Cost Bibliography, http://blogs.loc.gov/digitalpreservation/2012/06/a-digital-asset-sustainability-and-preservation-cost-bibliography/ [Accessed 31 Jan 2014]
[3] An expanded definition can be found in the appendix A.1—Terms and definitions

to apply the OAIS Reference Model referred to earlier to discussions in this document. This standard includes a functional model that describes a conceptual repository for long-term preservation of information and three roles that interact with the repository, namely Manager, Producer and Consumer (see Figure 1).

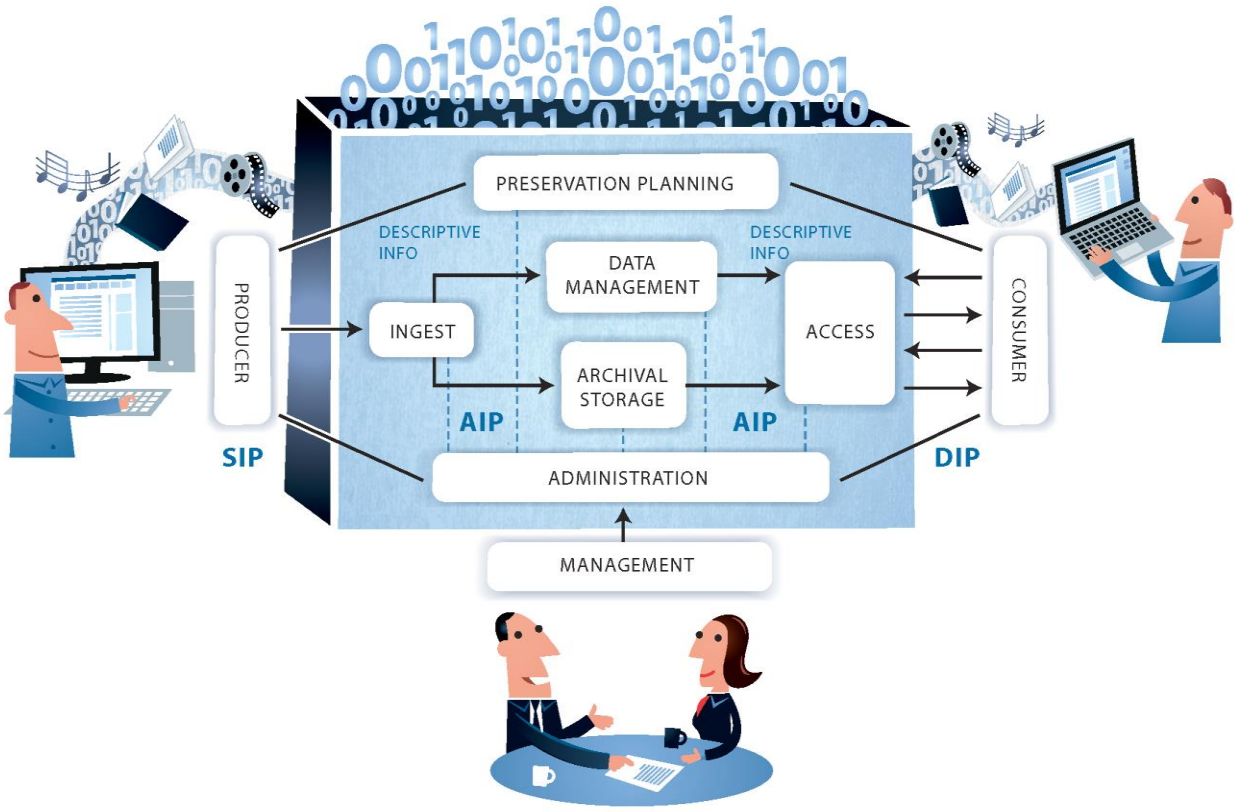

Figure 1—Representation of the OAIS functional entity model[4]

Thus, digital curation is interpreted as the activities that make up Ingest, Data Management, Archival Storage, Preservation Planning, Access, Administration, and Common Services. But digital curation extends beyond these core repository activities and also includes activities that may be seen as expressed by the OAIS roles. These activities encompass pre-repository activities, such as data creation and digitisation and pre-ingest (for example appraisal, selection and negotiation of agreements), and post-repository activities, for example re-use of data sets; and management activities, such as provision of budgets, policies and risk analysis. Furthermore, digital curation in this context does not only imply long-term storage, but also short and medium term storage (or, to be more precise, storage for as long as stakeholders are willing to pay for the services). On top of this, digital curation also covers information asset management systems and services with less strict specifications for archiving than those required by the OAIS standard. Whilst repository activities are well defined by the OAIS standard, the activities encompassed by the roles generally are not. However, activities involved in the interaction between the role of the Producer and the repository are described in detail by the "*Producer-Archive Interface Methodology Abstract Standard"* (PAIMAS)[CCSDS, 2004, (ISO 20652:2006)].

[4] Digital Bevaring, http://digitalbevaring.dk [Accessed 31 Jan 2014]

We now return to the last part of the stated task, namely "... *to produce a needs & gap analysis*". 'Needs' refers to the stakeholders' needs for financial information, in other words, their need for factual cost data and contextual information that defines these costs. The aim of the gap analysis is to reveal shortcomings in the capabilities of the models in relation to these user's needs. Gaps include not only deficiencies in the capabilities of individual models, but also gaps in cost and benefit modelling methods as a whole.

In addition this report recommends ways to bridge the identified gaps to utilise existing knowledge and facilitate the use and uptake of effective modelling methods. Finally, the report identifies future research opportunities.

1.2 Related work

This work aims to synthesize and utilise existing knowledge about how to determine the costs and benefits of curating digital assets. To a large extent the work presented in this report benefits from the recent work done by the Alliance [to enable] Permanent Access to the Records of Science in Europe Network (APARSEN) project[5] which also analysed existing cost models within the field of digital curation [APARSEN, D32.1-2, 2013]. However, the APARSEN project focuses on benchmarking cost models against the ISO 16363 standard "Audit and Certification of Trustworthy Digital Repositories" [CCSDS, 2012 (ISO 16363:2012)]. This is done by mapping the activity breakdown of the models to the activities framed by the standard, namely the activities within the three main sections "Organisational Infrastructure", "Digital Object Management" and "Infrastructure and Security Risk Management", as well as the sections' sub levels.

In contrast, wherever the 4C project focuses on analysing the models the mapping is made against the stakeholder's needs. This focus allows us to assess whether the models meet the *actual* needs of the stakeholders and to identify any existing gaps. Furthermore, this analysis not only focuses on the activity breakdown of the models, but also on how costs are broken down by resource type and time, what cost variables are included, and on various aspects of the models' functionalities, usability, and the like. Consequently, there are significant differences in the purpose and methods used to analyse the models.

The work done by the APARSEN project, however, provides a unique opportunity to draw upon results, experience and lessons learned by others. In particular the 4C project has benefited from studying the APARSEN parameters for analysis of cost models and some of their stakeholder consultation questions for evaluating the use of cost models. Both of these resources were used as inspiration, but were expanded and adjusted to fit to the purpose of this work. Another lesson learned from the APARSEN project was that cross testing models with cost data is extremely difficult due to the structural differences in the existing cost models. The attempt to cross test illustrated clearly that the models are not convertible.

The APARSEN report [APARSEN, D32.2, 2013, p. 16, 18, 21] point at gaps in existing models and make recommendations for future work. The points of particular interest for this work are:

Lack of activities (in relation to ISO 16363):

> *'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.'*

[5] APARSEN, http://www.alliancepermanentaccess.org [Accessed 31 Jan 2014]

Lack of clear terminology

> *'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.'*

Lack of interoperability between cost models

> *'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 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.'*

Apart from the work of APARSEN this work also draws upon the extensive bibliography made by the LIFE project [Watson, 2005], The Signal[6], the Open Planets Foundation (OPF) website[7] and in the 4C report [4C, D2.1, 2013], which lists relevant initiatives on cost modelling for digital curation.

[6] The Signal – Digital Preservation, A Digital Asset Sustainability and Preservation Cost Bibliography, http://blogs.loc.gov/digitalpreservation/2012/06/a-digital-asset-sustainability-and-preservation-cost-bibliography/ [Accessed 31 Jan 2014]
[7] Open Planets Foundation, Digital Preservation and Data Curation Costing and Cost Modelling, http://wiki.opf-labs.org/display/CDP/Home [Accessed 31 Jan 2014]

2 Characteristics of Cost and Benefit Models

In order to understand both the evaluation of cost and benefit models and the discussion of the models' characteristics, a basic terminology and a conceptual description of models and their components is important. The core characteristics are outlined here.

2.1 Cost models

We define a cost model for digital curation as a representation of the resources, such as capital and labour, used for digital curation activities. The use of these resources can be measured in different ways and by applying a price for their use they can be evaluated in monetary terms, their cost. Cost models can further be characterised by how they structure cost data, how they model variables, how they record cost data and by their usability. These four characteristics are further described below.

2.1.1 Cost structure—activity, resource and time

We use the term cost structure to refer to the way a model defines and divides costs in elements by the dimensions activity, resource and time.

2.1.1.1 Activity

Costs are typically divided according to what activity they account for. An activity can consist of one or more functions or processes that can be broken down to different levels of granularity.

Typically activities are structured in categories and different levels of sub categories. Many of the existing cost models use the OAIS standard [CCSDS, 2012 (ISO14721:2012)] as a point of reference for describing activities that incur costs. Note that the definition of digital curation and underlying digital curation activities are not universally accepted and understood, nor sufficiently detailed.

Furthermore, the measure of quality of curation activities is still unclear and difficult to account for. This represents an important challenge in cost and benefit modelling because there is a close relationship between the quality of the curation activities (and the resulting quality of the information assets) and the perceived benefits.

Some cost models include activity checklists that serve to check if the required digital curation activities are included.

2.1.1.2 Resource

Costs can also be divided according to the type of resource, capital or labour, they refer to. Capital cost can be further differentiated by type—examples include building space (server space, office space, and so on), equipment (servers, network, and the like), energy (for systems, cooling, et cetera) and materials (storage media, and so on). Labour costs can be differentiated by level of education (unskilled, skilled, 1st degree, Masters degree, Doctorate, and so on) and/or job functions (developer, metadata officer, et cetera).

Just as it can be difficult to segregate costs in activities it can be difficult to segregate costs within resources. The Transparent Approach to Costing (TRAC)[8], which is applied in Higher Education in Britain, has been suggested as a methodology for recording resource cost data [Beagrie et al., 2008, p.13].

Direct and indirect costs—variable and fixed costs

Direct costs are those associated with resources used for performing digital curation activities, such as costs of acquisition of storage media or the costs of staff employed to add metadata, where the amount of resources spent can be directly measured.

Indirect costs are those incurred by the usage of shared resources, such as general management and administration or common facilities and systems, where it has not been possible to distribute the cost on specific activities. Thus the lack of detailed measures of the use of shared resources often lead shared cost to be indirect cost.

Variable costs are costs, which vary directly with the amount of production, and they are therefore normally equal to direct costs.

Fixed costs are costs, which do not vary with the amount of production, and therefore they are normally equal to indirect costs.

For lack of a better measure of indirect costs—also called residual costs or overhead—they can sometimes be added to direct costs as a percentage of direct cost. In this case indirect costs are not directly equal to fixed cost. In general, given enough scale and time, no cost is really fixed.

2.1.1.3 Time

Costs can be divided by accounting periods to capture past cost (ex post) and/or future costs (ex ante). Records of past cost are used in accounting whereas estimations of future costs over certain time periods (such as months, quarters, years) are used for budgeting. Estimations of costs are often based on analogy, in other words on experience from similar activities and projections of historic cost data, for example derived from accounts.

The following time aspects are also referred to as financial or economic adjustments in some cost models.

One-time, periodic, recurring costs

Costs can be divided by time such as one-time costs, periodic (term) costs or recurring costs, depending on the time period. The term capital or investment cost is often used to denote a one-time cost incurred on the acquisition of equipment such as a storage system. The term periodic cost is used to indicate that the cost will incur at intervals. Recurring costs also known as running costs or operating costs include costs of the consumption of media, energy and labour[9].

Depreciation

Depending on the unit of time costs can be expressed as the depreciation (physical or through obsolescence) of assets. For example, the time in which a server becomes obsolete (one measure of the lifetime of a server) may be five years. With a 5-year time period the cost of using this resource may simply be its acquisition cost, whereas with a 1-year period the cost would be the depreciated acquisition cost (whether linear, exponential or other). In general, depreciation (for tangible assets) and amortization

[8] TRAC, http://www.jcpsg.ac.uk/guidance/ [Accessed 31 Jan 2014]
[9] Capital costs are often abbreviated as capex (capital expenditure), and operating costs as opex (operating expenditure)

(for intangible assets) are mechanisms for distributing capital costs over the estimated useful lifetime of an asset to indicate how much of an asset's value has been used.

Inflation, interest rates, discount rates

Other important time aspects of costs are the time value of money such as inflation (general price increases), individual price changes that are related to specific resources—such as storage media, energy, office space, computer scientist wages—and interest, which reflect economic growth and cost of capital.

Even though the cost of resources have in general been increasing, the cost of both capital and labour per unit of digital information assets has, due to technological innovation, been decreasing over the past decades, although at very different rates. Therefore, in order to calculate the present value of estimated future costs different discount rates are preferable. The present value is needed in order to compare different cost scenarios over time.

2.1.2 Cost variables

Cost models can also be characterised by the way they handle cost variables—the factors that influence the cost. The cost variables represent contextual information that is essential for understanding the assumptions underlying actual cost data and for comparing scenarios. An additional sub division to divide cost variables into service adjustments and economic adjustments has been proposed [Beagrie et al., KRDS, 2008, p. 20].

Service adjustments are defined as adjustments in relation to the assets, or the digital curation system and/or service. The former includes adjustments of the *quantity* of the assets, expressed as numbers of items and/or by data volume, and the *quality* of the assets, meaning the type and complexity of the assets, and their significant properties. Digital curation system and/or service adjustments relate to adjustments of the *quality* of systems and services, for example the degree of a repository's trustworthiness or the quality of an error handling procedure.

Economic adjustments include inflation (or deflation), depreciation, and interest (discount rates).

2.1.3 Recording cost data

Cost data can be recorded in various ways depending on what the data are needed for and the perspective of a cost analysis. Table 1 below shows a generic example of a cost data record. In this table the red text indicates cost parameters and blue text represents values. Typically, accounts and budgets only include selected cost parameters, for example investment (one-time, direct capital) costs, operating (recurring, direct, capital) costs, labour (recurring, direct, labour) costs and indirect costs. We use the term 'parameter' to express these various cost elements, for example "labour costs", "capital costs" and cost variables, such as "quantity of assets", "Salary level". Each parameter has values assigned which can be either numerical, Boolean, or an ordered list.

	Time								
	Period 1								Period 2…
	Resource (€)								**Resource**
Activity	One-time				Recurring				
	Direct		Indirect		Direct		Indirect		
	Capital	Labour	Capital	Labour	Capital	Labour	Capital	Labour	
A	100,000	50,000	1,000	1,000					
B									
C									
D …									

Table 1—Example of a generic cost template where the cost elements are recorded by activity, resource, and time.

2.1.4 Usability

Cost models can also be characterised by their usability. In general the usability of a cost model depends on how well it is designed, the accompanying documentation and the UX[10] design applied to the interface. This will determine how easy it is to understand, how easy it is to learn to use and operate.

Cost models can include a set of mathematical equations that convert resource data into cost data. In some cases the models are implemented in software, such as spreadsheets to simplify the application of these equations. Such tools can include pre-set parameters, functions and values to help guide users to best practices. Values may be assigned directly to a parameter or generated as the result of a function. As mentioned, some cost models also have pre-defined activity checklists.

2.2 Benefit models

A benefit model is defined as a representation that describes the benefits and value incurred by digital curation activities. Benefits are typically divided in financial benefits—benefits that can be expressed in monetary values (for example value generated from user fees or licenses)—and in non-financial benefits—benefits in the form of an organisations' increased trustworthiness (reputation) or reduced business risks.

2.3 Economic models

Economic models go beyond cost and benefit models and describe the economic processes around digital curation work; including the flow of resources (costs and revenues) within the economic lifecycle of digital information assets, and stakeholders (from the demand, supply and management side) interaction with this lifecycle.

Figure 2 shows an overview of the relation between economic models and cost and benefit models [4C, MS9, 2013, p. 41].

[10] UX—User Experience. A design process designed to facilitate the use of man machine interfaces.

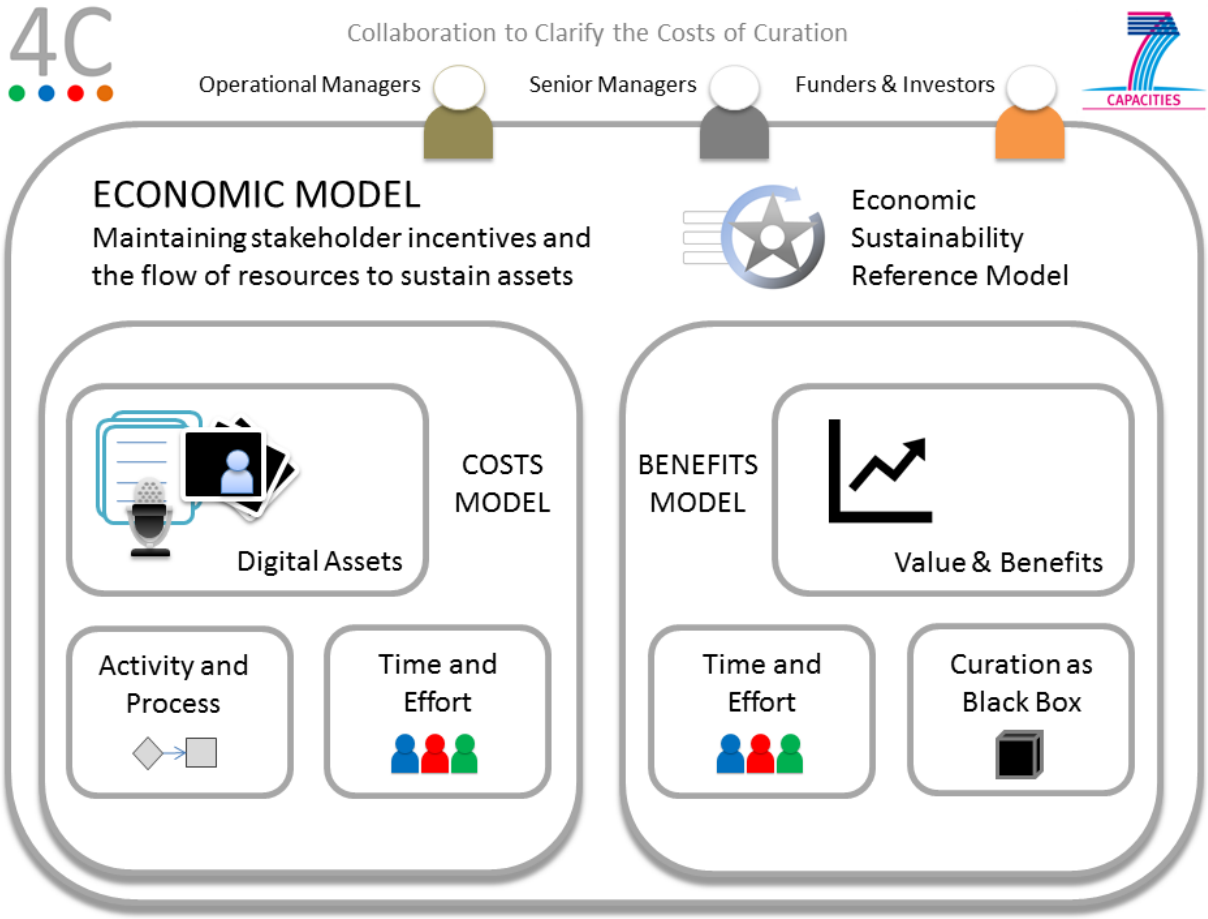

Figure 2—The nesting of Cost and Benefits modelling activities within the overarching framework of an economic model

3 Description of Existing Models

The first process undertaken in the evaluation of the existing models was to study recent work and research on the topic and identify the models relevant to the field of digital curation. This task was crucial as it forms the foundation of the subsequent work. The challenge of this task was to make sure all relevant models were considered and that the appropriate models were identified and included in the further analyses. The identified models differ vastly in focus, design, functionality, cost types and so on. This section gives an overview of the models and their capabilities.

3.1 Identification of models

This project had previously conducted a literature review of existing work and research that was used to provide a starting set of models [4C, D2.1, 2013]. In addition there was a call for further model nominations from within the community via the 4C website[11].

We selected models that cover the lifecycle of digital curation (fully or partly) and which model cost and/or benefits of curation activities. Models that were not published or where documentation and/or papers were not publicly available were not considered, as they would be impossible to analyse equitably.

Originally we had also included the Economic Sustainability Reference Model (ESRM) in the list of identified models[12]. Based on the findings and recommendations in the final report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access [BRTF, 2010], the ESRM breaks economically sustainable digital curation down into four primary components:

- the economic lifecycle
- the sustainability strategy
- economic risks and remedies, and
- key entities.

These provide a framework that assists in thinking through sustainability issues over the complete lifecycle for digital assets.

However, this model was taken out of the evaluation because it is an economic model that addresses sustainability of digital curation at a higher level than that of cost and benefit models. It does not describe an architecture, nor does it have an implementation. We therefore found that was inappropriate to evaluate it against requirements for cost and benefit models.

The ESRM

The ESRM is best understood as a strategic tool for planning and discussion aimed at executive and managerial rather than operational level staff. Its purpose is to provide a foundation for progress in the development of successful sustainability strategies for digital curation. It does this by organizing the problem space; providing a common reference point of concepts and vocabulary; and introducing a layer of abstraction that hides the complexities and idiosyncrasies of individual implementations and contexts, while at the same time embodying sufficient detail to support substantive discussions of shared issues.

[11] Call for Curation Cost Models, 4C blog post June 2013, http://www.4cproject.eu/news-and-comment/4c-blog/18-call-for-curation-cost-models-by-ulla-bogvad-kejser [Accessed 31 Jan 2014].
[12] Update on the state of the Economic Sustainability Reference Model, https://unsustainableideas.wordpress.com/2011/10/17/update-state-ref-model/ [Accessed 31 Jan 2014]

3.2 Overview of models

The identification process produced a list of 10 models that were considered relevant to the field of digital curation. These are listed in Table 2.

ID	Name	Acronym	Owner
1	Test bed Cost Model for Digital Preservation	T-CMDP	National Archives of the Netherlands
2	NASA Cost Estimation Tool	NASA-CET	National Aeronautics & Space Administration
3	LIFE^3 Costing Model	LIFE3	University College London and The British Library
4	Keeping Research Data Safe	KRDS	Charles Beagrie Limited
5	Cost Model for Digital Archiving	CMDA	Data Archiving and Networked Services (DANS)
6	Cost Model for Digital Preservation	CMDP	Danish National Archives and The Royal Library, DK
7	DP4lib Cost Model	DP4lib	German National Library
8	PrestoPRIME Cost Model for Digital Storage	PP-CMDS	The PrestoPRIME project
9	Total Cost of Preservation	CDL-TCP	California Digital Library
10	Economic Model of Long-Term Storage	EMLTS	Rosenthal, D.

Table 2—List of models identified as relevant to the field of digital curation

3.3 Model details

A detailed description of each of the 10 models is provided below. The model descriptions are structured in a schema allowing the same generic characteristics to be described for each model. These characteristics are:

- ID—our internally generated Identification number (see Table 2 above)
- Model name—name and acronym (we have assigned acronyms to all models, using existing ones where available)
- Creators and funding—the creators and/or owners of the model and, where known, the funders of the model
- Status—information regarding the current status of the mode including such matters as the current version (and version date), future developments planned (if any), and the like
- Purpose—the underlying purpose (mission) of the model
- Information Assets—the type of curation assets the model is designed to deal with
- Activities—the activities the model identifies and/or models
- Resources—the types of resources applied by the model, for instance, capital cost, labour cost, and so on.
- Time—time period over which the model is intended to model the curation cost
- Variables—the type of variables that the model handles, such as number and type of assets
- Type of tool—the generic tool type—for example analysis and/or estimate—and in some cases information about the underlying technology (Microsoft Excel spreadsheet for instance)
- Availability of tools—the existence (or absence) and type of user accessible tools along with details of where/how to access them.
- References—to further reading.

3.3.1 Test bed Cost Model for Digital Preservation, T-CMDP

T-CMDP gives a list of cost indicators, which influence the total cost of preservation. It also includes a computational model for calculating the total cost of preservation.

The model defines five high level cost indicators:

- cost of a digital repository and preservation system
- personnel costs
- cost of developing or acquiring software and strategies
- cost of performing preservation actions, and
- other costs

These cost indicators are used to estimate the total cost of preservation and expresses it as money and time. The model is based on OAIS terminology. It examines migration on request and migration to XML (preservation format) and emulation using the Universal Virtual Computer (UVC).

Included in the model is the cost of the digital archival system (a digital depot or repository) and functionality for the long term preservation of digital records; personnel costs, such as the cost of the development or procurement of software and methods for the preservation of digital records; the cost of the actual storage of digital records; and other factors that exert an influence on the total cost.

The tool is a structured with: a costs basis (labour, capital); operational cost: labour cost per preservation activity per asset type (email, text, spreadsheet, database); activities: Acquire and appraise, metadata, repair, develop preservation approach (gather req. and develop approach), preserve and evaluate.

The model is straightforward and easy to use. However, it’s out of date and does not breakdown activities in sufficient detail. It could be updated to capture greater detail but the LIFE work probably has already done this so there isn't much benefit in revising this model further.

Property	Description
ID	1
Name	Test bed Cost Model For Digital Preservation (T-CMDP)
Creator and funding	Developed by the National Archives of the Netherlands as a part of the digital preservation test bed project.
Status	The latest versions of the model and the computational model spreadsheet (Version 1.0) are from 2005; version 1.2 30-aug-2005 for the cost model tool.
Purpose	The purpose is to estimate the costs of long-term preservation and compare the costs involved in applying different preservation approaches.
Information assets	Texts, email, spreadsheets, databases
Activities	Ingest, Archival Storage, Data Management, Administration, Preservation Planning (normalisation at ingest and migration after 20 years)
Resources	Capital cost, labour cost (6 types); fixed cost, operational cost.
Time	Present, future
Variables	Labour salaries (6 types), capital cost (building space, hardware and software (clients, servers, databases, storage, "archive system"), migration frequency, number of assets
Type of tool	Analysis tool, implemented in spreadsheet
Availability of tools	The spreadsheet is no longer available at the National Archives of the Netherlands, but can be found on the Internet Archive (https://archive.org/)
References	Slats, J. and Verdegem, R. (2005) "Cost Model for Digital Preservation", Proceedings of the IVth triennieal conference, DLM Forum, Archive, Records and Information Management in Europe: http://dlmforum.typepad.com/Paper_RemcoVerdegem_and_JS_CostModelfordigitalpreservation.pdf

Table 3—T-CMDP characteristics

3.3.2 NASA Cost Estimating Tool, NASA-CET

The NASA-CET (CET) is designed to generate life-cycle cost estimates for implementing, operating and maintaining a science data system. It employs the cost estimation by analogy approach, using information about existing data activities as the basis for estimating life cycle-costs for user-defined activities. It is intended for use on science data, but can be applied to other institutions.

The CET divides the life-cycle into a set of functional areas that cover different generic areas of costs. These compose the reference model behind the tool. The functional areas include: ingest, processing, documentation, archive, access and distribution, user support, sustaining engineering, engineering support, technical coordination, implementation and facility /infrastructure. Based on the functional areas users can describe activities and the CET then estimates the costs for the user-described activity and outputs the results in spreadsheet and graphical formats. The effort is expressed in Full Time Equivalents (FTEs), which is equivalent to that of a single person working full time for a year. With this model and its tool a user can enter their own data set for a new similar activity, estimate its cost, and review it.

The CET uses regression to develop the coefficients for a set of seven trial relationships of FTE to workload parameter for each of the selected workload parameters.

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

For the first three relationships, the CET uses single parameter regression of Y's on X's (effort on workload), and for the last four, two-parameter multiple regression; for example, for the quadratic case the two parameters are X and X^2.

The estimate is by analogy, in other words the data sets come from similar NASA missions with similar activities. A tool has been made to support the import of the data sets to the comparable databases.

The model has a data activity reference model. OAIS is used as a reference for the model (a mapping has been made). Note that OAIS preservation planning and migration activities are not explicitly included in the model and its tool. CET does not directly address long-term archiving; it would need extensions to be able to do that.

The NASA-CET model covers a lot of detailed information for in-project costs so could be very useful for Principal Investigators (PIs) wanting to assess Research Data Management (RDM) and sharing costs for new research projects. However, the learning curve is probably too steep and the amount of time needed to capture and model the activities to enter into the tool may require too much for most researchers to want to use it. The level of complexity in the tool may be usual for the space sciences but for most other disciplines it is far too complicated. If it were simplified greatly though, it could be very useful as a project costing tool.

Property	Description
ID	2
Name	The NASA Cost Estimation Tool (NASA-CET)
Creator and funding	Developed for NASA (National Aeronautics and Space Administration) by SGT (Stinger Ghaffarian Technologies, Inc.)
Status	The first version of CET (Version 1) was published in 2004. The newest version available (Version 2.4) is from September 2008.
Purpose	Estimating life cycle costs for ground data centres activates to improve budgets for NASA missions.
Information assets	Space data—typicaly multidimensional data sets
Activity	Ingest, Data Management, Archival Storage, Access, Administration
Resource	Capital cost (system purchase, maintenance, commercial of the shelf (COTS), source lines of code (SLOC), facility, media), labour cost (5 levels)
Time	Past, present and future—for a 7 to 10 year time scale, reflecting normal data processing time period for missions.
Variables	94 distinct descriptors, such as staff salaries, system purchase cost, commercial off the shelf software license, archive media, inflation, volume, automation level.
Type of tool	Analysis, estimation and review of estimation. Implemented in MS Excel spreadsheet using visual basic (VBA).
Availability of tools	The CET tool, users guide, technical description etc. is available for download at: http://opensource.gsfc.nasa.gov/projects/CET/index.php
References	Fontaine, K., Hunolt, G., Booth, A. and Banks, M., Observations on Cost Modeling and Performance Measurement of Long Term Archives, in PV2007 Conference Proceedings, 2007: http://www.pv2007.dlr.de/Papers/Fontaine_CostModelObservations.pdf Hunolt, G., Booth, B. and Banks, M., Cost Estimation Toolkit (CET) Users' Guide", Version 2.4, 2008, (available as part of the CET software package: http://opensource.gsfc.nasa.gov/projects/CET/index.php)

Table 4—NASA-CET characteristics

3.3.3 LIFE^3 Costing Model, LIFE3

LIFE3 provides a methodology to model the digital lifecycle and a tool to calculate the predicted costs of preserving digital information. The model was developed in the context of libraries and Higher Education/Universities but it can be applied to other cultural heritage institutions of any size.

LIFE3 is based on the OAIS model (and could be even closer to the OAIS model according to the LIFE2 evaluation). It incorporates the digital preservation costing model (GPM v1.2), developed during LIFE 1 and 2.

LIFE3 estimates the complete lifecycle cost as the cost of the lifecycle stages:

- Creation
- Acquisition
- Ingest
- Metadata Creation
- Bit-stream Preservation
- Content Preservation and Access

Each lifecycle stage has several lifecycle elements which in turn have sub-elements.

Regarding 'Content Preservation' three strategies of migration are possible: 'do nothing' (default), 'migrate on ingest', and 'migrate periodically'. The cost of migration is based on the complexity of the format.

The tool requires the user to enter information in only five fields. This information is used to pre-populate the model with data averaged from relevant case studies where it is available, and the user is immediately presented with a cost estimate on the output page.

The five required types of information is Start year/End Year, Category (asset type), Source (creating costs), Number of items for each year, Organisation Size.

The tool supports the following type of assets: Web sites, E-journals, Printed items (digitised), Sound recordings (digital or analogue), Research documents (MS Office, PDFs, small DB).

The total estimated lifecycle cost is calculated for each stage and element. The cost can be discounted. Migration costs are by default zero due to the default migration strategy of 'do nothing'. Changing this to migrate periodically and adding the required information on frequency and cost per item shows that the migration cost is the major cost in the model (regardless of a few calculations error in the current version of the tool).

The LIFE3 tool covers most aspects well from an archival point of view, but could benefit from a wider array of pre-ingest activity. Users would probably benefit from a quick user guide. The pre-populated spreadsheet developed by HATII to test the LIFE tool for pre-ingest activity might be worth updating and trialling more widely[13].

[13] DCC, Piloting the LIFE costs Tool in UK HEIs, http://www.dcc.ac.uk/projects/life [Accessed 31 Jan 2014]

Property	Description
ID	3
Name	LIFE^3 Costing Model (LIFE3)
Creator and funding	Developed by University College London (UCL) and British Library (BL) and funded by Jisc and Research Information Network (RIN)
Status	The LIFE3 project ended in 2010
Purpose	To improve the ability of organisations to plan and manage the preservation of digital assets by giving a content neutral view of the digital lifecycle from the perspective of the preserving organisation. In LIFE1 and LIFE2 the purpose was to estimate the life-cycle cost of preservation activities to aid decision making and budgeting
Information assets	Websites, e-journals, digitized newspapers, sound, word processing documents, small databases.
Activities	Production (creation/digitisation), Pre-ingest (acquisition) Ingest, Data Management, Archival Storage, Preservation Planning, Administration
Resources	Capital (storage media), labour (5 levels)
Time	Past, present, future up to 100 years
Variables	Retention period, number of items, types of items, migration (strategy, frequency, coverage, automation rate), organisation size
Type of tool	Analysis, estimate. The tool is implemented in a MS Excel spreadsheet, and in a prototype web version.
Availability of tools	Documentation and a predicting costing tool is available for download at: http://www.life.ac.uk/ Prototype web tool at: http://lifedev.hatii.arts.gla.ac.uk/
References	Hole, B., Wheatley, P., Lin, L., McCann, P. and Aitken, B. The Life3 Predictive Costing Tool for Digital Collections, in New Review of Information Networking, Volume 15, Issue 2, 2010: http://www.tandfonline.com/doi/abs/10.1080/13614576.2010.526014 Hole, B., Lin, L., McCann, P., and Wheatley, P., LIFE3: A Predictive Costing Tool for Digital Collections, in Proceedings of iPRES 2010, 7th International Conference on Preservation of Digital Objects, Austria, 2010, http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/hole-64.pdf Watson, J., The LIFE project research review – Mapping the landscape, riding a life cycle, 2005, http://discovery.ucl.ac.uk/1856/1/review.pdf Wheatley, P and Hole, B., LIFE3: Predicting Long Term Digital Preservation Costs, In iPRES2009, September 2009: http://www.escholarship.org/uc/item/23b3225n

Table 5—LIFE3 characteristics

3.3.4 Keeping Research Data Safe, KRDS

The KRDS models aims to give understanding of long-term preservation costs for research data and support cost benefit-analyses for justifying and sustaining major investments in digital repositories and curation. It is developed in the context of universities but is widely applicable.

The KRDS cost model is activity based and builds on OAIS terminology. It divides the digital curation life-cycle into phases consisting of activities and sub activities. Each of which represents a cost variable. The model is a framework and does not include a cost predicting tool.

The KRDS Benefits Analysis Toolkit includes the 'KRDS Benefits Framework' and the 'Value Chain and Benefits Impact Tools'. Each tool consists of a more detailed guide and worksheets. The combined Toolkit provides a very flexible set of tools, worksheets, and lists of examples of generic benefits and potential metrics.

The Benefits Framework Tool is for identifying, assessing, and communicating the benefits from investing resources in the curation and long-term preservation of research data. The benefits in this tool are divided into three dimensions which are all further divided into two categories. The first dimension consists of direct and indirect benefits; the second one is divided into near-term and long-term benefits while the third one includes internal and external benefits[14].

The more advanced Value Chain and Benefits Impact Analysis Tool is designed to be used for longer-term and intensive activities such as evaluation and strategic planning. The impact component of the tool helps identify potential quantitative metrics or qualitative indicators for the value of the benefits identified.

The activity model concepts cover all aspects of the lifecycle in the model framework but users need to develop their own calculations based on guidance in the user guide. This is not a bad thing as most organisations will want the model to reflect their own environment but some default data could be useful to avoid users having to develop their own formulas entirely from scratch.

[14] There is also a KRDS Stakeholder analysis version of the benefits framework available, for an example see "Benefits from Research Data Management in Universities for Industry and Not-for-Profit Research Partners", http://opus.bath.ac.uk/32509 [Accessed 31 Jan 20114].

Property	Description
ID	4
Name	Keeping Research Data Safe (KRDS)
Creator and funding	The KRDS project is funded by JISC and conducted by a partnership of the following institutions: Charles Beagrie Ltd, OCLC Research, the UK Data Archive, the Archaeology Data Service, the University of London Computer Centre, and the universities of Cambridge, King's College London, Oxford and Southampton.
Status	The KRDS project ended in 2010, though there has been some follow-up activity. The latest documentation dates from July 2011.
Purpose	Support for efficient digital repositories and data curation.
Information assets	Research data
Activities	Production, Pre-ingest (pre-archive phase), Ingest, Data Management, Archival Storage, Preservation Planning, Administration, Access (archive phase).
Resources	Capital costs (equipment costs, travel, consumables, estate costs), labour costs; indirect costs, outsourcing
Time	Past, present and future—medium to long term
Variables	Collection levels, preservation aims, number of depositors; number, mode and frequency of deposits; number, complexity and type of file formats
Type of tool	Analysis
Availability of tools	Tools, documentation, users guide etc. is available for download at: http://www.beagrie.com/krds.php
References	Charles Beagrie Limited, User guide for keeping research data safe. Assessing costs/benefits of research data management, preservation and re-use, Version 2.0. Copyright HEFCE 2010 and 2011: http://www.beagrie.com/KeepingResearchDataSafe_UserGuide_v2.pdf Beagrie, N. and Pink, C., 2012. Benefits from Research Data Management in Universities for Industry and Not-for-Profit Research Partners. Charles Beagrie Ltd and University of Bath. http://opus.bath.ac.uk/32509/ Beagrie, N., Chruszcz, J. and Lavoie, B. Keeping Research Data Safe. A Cost Model and Guidance for UK Universities, Copyright HEFCE 2008, http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf Beagrie, N., B. Lavoie, and M. Woollard, Keeping Research Data Safe 2, Final Report, Charles Beagrie Ltd., 2010, http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf Beagrie, KRDS/I2S2 Digital Preservation Benefit Analysis Tools Project, http://beagrie.com/krds-i2s2.php

Table 6—KRDS characteristics

3.3.5 Cost Model for Digital Archiving, CMDA

The model is an activity-based costing model for long-term preservation and dissemination of digital research data sets. Its purpose is to assist the management in achieving economic sustainability for data archiving.

The project behind CMDA started with a narrowly scoped problem focusing on creating and testing a model which would estimate the costs of research datasets preserved by a trusted digital repository. In the process, the scope of the project expanded into looking for a way to express the true value of the organisation to its stakeholders.

The CMDA model is based on the OAIS Reference Model. It identifies activities and a set of costing components of each activity. These are:

- Activities
- resource pools
- resource cost drivers
- activity cost drivers, and
- cost objects

The model also uses a matrix for raking complexity of datasets to take the influence of varying data complexity into account when estimating costs.

Based on these factors the model estimates costs per dataset. Costs are expressed as monetary costs specified for datasets in the unit ‘euros per dataset’.

The model is developed based on DANS’s curation of 14,000 datasets with a total size of 1.5 TB and 10 other datasets with a total size of 20TB. The data sets for archaeology are much more costly than social sciences or humanity due to more variety and complexity in data formats (databases, images, geodata, cad and so on). The activities “preservation” and “development of the archival system” are the most cost intensive.

CMDA uses a balanced scorecard (BSC) which was designed as a management system that enables organizations to clarify their vision and strategy and translate them into action. It also acts as a measurement system and communication tool.

The BSC has been divided into four perspectives:

- Supporters
- Internal Processes
- Customers
- Innovation and Growth

The tool also defines Success Factors which describe the strategic objectives of the organisation. They are described further by a set of Performance Indicators which are an indication of how it shall be known that the outcome has come to pass. Success Factors and Performance Indicators can be used to connect measured benefits to activities and costs.

The model’s activities and cost drivers assume that the organisation has the 'philosophy of a trusted digital repository (TDR) compliant with the 16 guidelines listed in the Data Seal of Approval (DSA)'. As such, the model would likely work quite well for other TDRs and also for other types of repositories.

Property	Description
ID	5
Name	Cost Model for Digital Archiving (CMDA)
Creator and funding	Developed by Data Archiving Networked Services (DANS) of the Netherlands
Status	The project creating CMDA has ended.
Purpose	Estimate costs in order to assist management in achieving economic sustainability for data archiving
Information assets	Research datasets from social sciences, humanities and archaeology
Activities	Ingest, Archival Storage, Preservation Planning, Data Management, Administration
Resources	Capital (office, minimal IT equipment), labour (general, archivists, ICTa, ICTb)
Time	Past and present
Variables	No. of files, metadata quality, complexity, type of personnel, quality standards applied
Type of tool	Accounting, analysis
Availability of tools	None. There is a description of the model at the webpage of DANS: http://www.dans.knaw.nl/en/content/categorieen/projecten/costs-digital-archiving-vol-2
References	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" in 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

Table 7—CMDA characteristics

3.3.6 Cost Model for Digital Preservation, CMDP

CMDP is a model for costing preservation of digital materials. It includes a tool that calculates present and future costs of cultural heritage institutions' digital collections based on various user inputs, such as the amount and type of data.

The model is intended for national cultural heritage institutions and was developed as a collaboration between a library and an archive. CMDP is activity based and adheres to the OAIS standard. It tries to estimate what OAIS activities and sub-activities are relevant for cost. The CMDP tool estimates costs and expresses the costs as monetary costs and/or person weeks.

CMDP was inspired by LIFE1 and developed simultaneously with LIFE2 and LIFE3 with meetings between LIFE and CMDP.

It is based on migration, normalisation at ingest as default and recurring migration. The experience of normalisation and migration at the Danish National Archives is used as the basis for estimating the cost.

CMDP is being developed iteratively and is currently missing a few OAIS entities. CMDP does not yet cover access, and only has very little financial adjustments such as depreciation. The model might be very useful if combined with KRDS approach or a simplified NASA-CET approach.

Property	Description
ID	6
Name	Cost Model for Digital Preservation (CMDP)
Creator and funding	Developed by the Danish National Archives (DNA) and the Royal Danish Library in Denmark (KBDK) and funded by the Danish Ministry of Culture.
Status	The latest version of the CMDP tool is from 2012.
Purpose	The purpose is to increase cost effectiveness of digital preservation activities and to provide a basis for comparing and estimating future cost requirements for digital preservation.
Information assets	Various types of office documents (unformatted/formatted text, spreadsheet, graphic, sound, video, hypertext, geodata, e-mail) and databases.
Activities	Pre-Ingest, Ingest, Archival Storage, Preservation Planning, Data Management (partially) Administration (partially)
Resource	Capital (storage equipment, migration equipment), labour (3 levels)
Time	Present, future (25 years)
Variables	Types of information assets (source/production format, destination/preservation format), volume per assets per year, type of storage system, number of copies of each information asset, salary-level, person-hours per week.
Type of tool	Analysis, prediction, implemented in a spreadsheet with modest use of VBA.
Availability of tools	The CMDP tool and documentation is available for download at: http://www.costmodelfordigitalpreservation.dk
References	Kejser, U.B., Nielsen, A. B. and Thirifays, A., Modelling the Costs of Preserving Digital Assets, in Proceedings of the UNESCO Memory of the World Conference, Vancouver, Canada, 2012: http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/mow/VC_Kejser_et_al_27_B_1350.pdf Kejser, U.B., A.B. Nielsen, and A. Thirifays, Cost Model for Digital Preservation: Cost of Digital Migration, The International Journal of Digital Curation 6, no. 1, 2011, 255-267, doi:10.2218/ijdc.v6i1.186: http://www.ijdc.net/index.php/ijdc/article/view/177 Kejser, U.B, Nielsen, A.B., Thirifays, A., Cost Aspects of Ingest and Normalization, Proceedings of the iPRES2011 Conference, November 1-4, 2011, Singapore, 1-10 Nielsen, A.B., Thirifays, A. and Kejser, U.B, Costs of Archival Storage, in Proceedings of the Archiving 2012 Conference, 2012, 205-210: http://www.imaging.org/IST/store/epub.cfm?abstrid=45307

Table 8—CMDP characteristics

3.3.7 DP4lib Cost Model, DP4lib

The DP4lib model is a cost-by-service model for long-term preservation of digital assets. It gives a framework in which all known costs can be recorded and allocated to a specific service. The model is developed by a library but covers any sector and can be used with different levels of detail.

The DP4lib model builds on the OAIS model. It categorises all costs into cost types which are further divided into cost elements. In estimating the costs all major cost elements—direct and indirect—are identified and attributed to the service that 'caused' them.

The model is accompanied by a tool for a specific implementation of the model, which can be used for calculating annual costs for long-term preservation expressed as money per year.

The model focuses on costing a service, and allocating (recording) the cost used for that service. Due to traditional accounting principles the cost is categorised as the types: hardware costs, software costs, people costs, accommodation costs, external service costs and transfer costs which are further broken down in cost elements (for example storage, OS, payroll, offices, recovery, internal charges). The model tries to allocate the many indirect costs to the services, so called absorbing.

The main services Ingest, Curation (Archival Storage, Data Management, Preservation Planning, Administration and Management) and Access are further divided in sub-activities. The model does not include the cost of migration. Apparently no migration (normalisation) at ingest is performed.

In order to distribute the many indirect costs a cost distribution key system is used. Equipment is depreciated based on acquisition cost and estimated life time.

Using the model on the data from the German National Library shows that the equipment is mainly storage. The cost of software is mainly annual license fee. The external services cover a major part of the archival information system, including operation and support of the DIAS system. Ingest, Curation and Access each account for about 1/3 of the cost. Internal labour accounts for about 4/10, internal hardware for 2/10, external services for 2/10, software 1/10 and buildings 1/10.

Property	Description
ID	7
Name	DP4lib Cost Model (DP4lib)
Creator and funding	Developed by the German National Library (DNB) and funded by the German Research Foundation (DFG).
Status	The latest version of the DP4lib model and tool is from 2012. Validation of the model is taking place in 2013.
Purpose	The purpose is to support estimating the costs for budgeting, accounting or charging.
Information assets	Digital documents, digitised books
Activities	Ingest, Archival Storage, Data Management, Preservation Planning, Administration, Management, Access
Resource	Capital (hardware, software, external service), labour; direct and indirect costs
Time	Past, present
Variables	Hardware, software, people, accommodation, external service, transfer - broken further down; linear depreciation of capital (acquisitions); keys for distributing indirect cost
Type of tool	Recording costs, analysis
Availability of tools	The documentation and description of the model is available online at: http://dp4lib.langzeitarchivierung.de/index_downloads.php.de (The tool is only available in German, and not directly available on the web site)
References	Report on the DP4lib cost model, A Cost Model for a Long-Term Preservation Service, 2012: http://aparsen.digitalpreservation.eu/pub/Main/CostModels/DP4lib-Cost-By-Service-CostModel.docx DP4lib Kostenmodell für Langzeitarchivierung, http://dp4lib.langzeitarchivierung.de/downloads/DP4lib-Kostenmodell_eines_LZA-Dienstes_v1.0.pdf

Table 9—DP4lib characteristics

3.3.8 PrestoPRIME Cost Model for Digital Storage, PP-CMDS

The PrestoPRIME model and tools were developed within an EC project and provide information for forecasting the costs of long-term preservation of mass digitised AV materials and comparing the costs of different preservation solutions.

The tools can be used for modelling and simulating the costs and risks of using IT storage for the long-term archiving of file-based audio visual assets. The objective is not to provide exhaustively detailed models or highly accurate costs but to provide simple results that are meaningful and useful in specific contexts, for example when selecting storage strategies or supporting day-to-day decision making. It accounts for cost as monetary expenses and risks of loss of assets.

The total annual cost of managed storage per copy is approximately 10 times today's cost of commodity storage media. The lifetime cost of managed storage per copy is approximately 4 times the current total annual cost of managed storage.

The cost levels incorporated in the model are based on experience from storage providers and a few large scale institutions. They conclude that cost-effective long-term storage at any scale requires automation through mass storage devices (disk servers and tape robots). The short obsolescence times for the components of these systems, including media, require the use of continual migration and active management.

The long-term storage planning tool allows the user to define storage media types (called storage systems), combine them in storage configurations and use them with collections. Media types can be added with storage migration period, running storage cost per volume unit per year with increasing or decreasing cost. The same can be done for running access cost. Corruption rates, latent and on access can be entered. Storage configurations are combinations of storage media types with given access share and frequency—and scrubbing. Collections can be added with average file size, length of cost/loss projection, amount of file and rate of change. The final plan consisting of a collection and a storage configuration can then be evaluated, with the result showing as three graphs over time, showing risk and loss, costs and corruption and loss.

The Storage simulation tool named iModel is a desktop GUI application, running on Java. It simulates the work environment of an archive manager by requesting that they take action as time goes by, requiring that they allocate resources where most needed. Actions can be file ingest, integrity checks, file repair using an alternative file copy, file access and so on. It has the same elements as the long-term planning tool, and can also simulate catastrophic events, such a losing all data on one set of media.

Tools for higher level preservation modelling, for example file format migration, digitisation and cataloguing that apparently were under development have not been released.

Property	Description
ID	8
Name	PrestoPRIME Cost Model for Digital Storage (PP-CMDS)
Creator and funding	Developed within the European Commission 7th Framework Programme for Research and Technological Development
Status	The prestoPRIME project ended in 2012.
Purpose	Practical solutions for the long-term preservation of digital media objects. Cost of storage and access.
Information assets	Audio visual material, film
Activities	Archival Storage, Access, Administration
Resources	Total cost, no specification of capital cost or labour cost, based on experience from storage providers.
Time	Present, future – up to 25 years
Variables	Data volume, migration frequency, latent and access corruption rates, no. of copies, storage systems, current costs, half-life for cost, access costs
Type of tool	The long-term planning tool is a storage planning tool, in the form of a web application running on Firefox and Chrome. The Storage simulation tool named iModel is a storage planning tool in the form of a desktop GUI application, running on Java.
Availability of tools	The storage simulation tool called iModel is available for download at: http://prestoprime.it-innovation.soton.ac.uk/imodel/download/ A simple storage planning tool is available for online use at: http://PrestoPRIME.it-innovation.soton.ac.uk/planning-tool/accounts/login?next=/planning-tool/
References	Addis, M. and Jacyno, M., Tools for modelling and simulating migration based preservation, PrestoPRIME Consortium, 2010, https://prestoprimews.ina.fr/public/deliverables/PP_WP2_D2.1.2_PreservationModellingTools_R0_v1.00.pdf Westerhof, H., Ubois, J. and Snyders, M., Financial Models and Calculation Mechanisms, PrestoPRIME Consortium, 2011, https://prestoprimews.ina.fr/public/deliverables/PP_WP6_D6.3.1_FM_calculation_R0_v1.01.pdf

Table 10—PP-CMDS characteristics

3.3.9 Total Cost of Preservation, CDL-TCP

The Total Cost of Preservation (TCP) model from the California Digital Library (CDL) is an analytical framework for modelling, assessing and accounting for the full economic costs of preservation. It relies on a number of fundamental abstractions and assumptions about preservation activities.

The model defines 10 high-level categories that cover all digital curation activities. The categories are based on the OAIS Model, but some of the concepts and terminology are modified to broaden applicability and facilitate understanding by non-specialists. Each category represents a cost component in the TCP pricing model and based on these components the pricing models calculate costs and express it as monetary expenses.

The 10 categories are:

- content owners
- submission streams
- preservation system (ingest, data management, access)
- servers
- storage
- consumers
- preservation planning
- interventions (e.g. migrations)
- administration
- management

The framework provides two price models that account for two different types of funding: Pay-as-you-go (PAYG) and Paid-UP (front) (PUP).

The Total Cost of Preservation model relies on a number of fundamental abstractions and assumptions about preservation activities. The cost associated with content creation or acquisition, reformatting, packaging, submission, and so on are excluded from the model. The cost of supporting owners in making use of the preservation System functions (sheet W3 in the spreadsheet tool) is included. Costs are nominal, based on generic instances of activities. However, preservation actions are assumed to be substantially automated, and the main costs are therefore acquisition and deployment of the software which is assumed to be independent of the number of objects.

System, Administration and Management are considered to be fixed costs—independent of the number of objects—whereas other costs are considered variable, in general proportional, using unit costs and number of units, such as the number of unique submission streams and the unit cost of a stream. The cost is assumed to be paid by the owners and the preservation service provider.

Establishing who pays for the cost is not essential for using the model, you can estimate the cost without that information. It has been incorporated as a factor to make the model more useful for CDL.

In the case of Paid-Up it is assumed that the archive institution (preservation service provider) can carry forward surpluses across fiscal year boundaries and reinvest them at market rates. (Note: For many public sector institutions this is not possible, therefore they are forced to use PAYG). Normal discounted cash flow analysis (DCF) is used. The model creator is aware of the shortcomings of DCF regarding fluctuating interest rates and the strong bias for the short term for the time value of money.

Property	Description
ID	9
Name	Total Cost of Preservation (CDL-TCP)
Creator and funding	The model was (and is being) developed by the California Digital Library (CDL), UC Curation Center (UC3) under a Creative Commons Attribution-Sharealike 3.0 license
Status	The latest version of the TCP pricing model tool and whitepaper is rev 2.1 from 2013-08-05
Purpose	Modelling the full economic costs of preservation, the "total cost of preservation" (TCP) over time in order to sustain long-term preservation efforts—effective and affordable curation management. UC3 itself needs a TCP model in order to move many of its core service offerings to a cost recovery operational basis.
Information assets	Any kind of digital asset—the model uses a generic, abstract level
Activities	Ingest, Data Management, Archival Storage, Preservation Planning, Access, Administration, Management
Resources	Total cost; in the tool total cost is refined into subsidiary costs such as capital cost, labour cost; operational cost; one-time, term or annual costs (called scope), fixed cost or marginal cost (proportional cost). Term costs are annualized over their lifespan and adjusted for inflation.
Time	Present, future—10 year scope
Variables	More than 100. For example, for "Migration" there are unit costs for: refreshment, replication, repackaging, transformation. For "Staff" there are 12 kinds of roles with salaries, FTE day rates.
Type of tool	Analysis tool, implemented as a MS Excel spreadsheet
Availability of tools	The tool is available for download at: https://wiki.ucop.edu/display/Curation/Cost+Modeling
References	California Digital Library, Total Cost of Preservation (TCP), Whitepaper, 2013,: https://wiki.ucop.edu/download/attachments/163610649/TCP-cost-price-modeling-for-sustainable-services-v2_1.pdf?version=4&modificationDate=1375721821000

Table 11—CDL-TCP characteristics

3.3.10 Economic Model for Long-Term Storage, EMLTS

The purpose of the model is to predict and compare the cost of long-term storage over time. The model predicts the costs of long-term storage for a data unit over a 100-year period. It covers the cost of capital of storage, and does not account for labour costs, presumably because labour cost are expected to be minimal for large amounts of data.

The model is not specific to any type of material as it focuses on storage. The model is a Monte Carlo simulation of the economic history of a unit of stored data. It has four high level components that model the costs of long-term storage:

- Yield Curves
- Loans
- Assets
- Technologies

Based on these four parameters the costs of storing one data unit is predicted and expressed as monetary costs.

This way of modelling storage cost is somewhat similar to known ways of continuously replacing equipment over time, but with a focus on uncertainty, both on the future interest rate used for discounting and on the future decrease in storage costs. Neither is truly exponential.

The model is based on up-front payment for long term (100 years) storage, not pay-as-you-go. The model uses discounted cash flow (DCF) to compare cost over time multiplying the interest rate by a short term factor and adds a planning horizon in years, which has to be paid off regardless of service life.

All storage technologies have a purchase cost, running cost, migration cost and a service time. New technologies arrive each year, and a technology incurs a purchase loan for the purchase cost and migration cost related to the previous technology, with a term equal the service life. Cost levels are based on experience from storage providers and a few large scale institutions.

After a few thousands run a Monte Carlo simulation can normally show what combinations fail, and what are durable.

Property	Description
ID	10
Name	Economic Model of Long-term Storage (EMLTS)
Creator and funding	Developed by David Rosenthal
Status	The latest blog post on the subject is from 2011
Purpose	To predict and compare the cost of long-term storage over time
Information assets	Any kind of digital asset, focus on long-term storage only, binary volume
Activities	Archival Storage, Administration
Resources	Total cost, no specification of capital cost or labour cost, based on experience from storage providers.
Time	Future—up to 100 years
Variables	Uses four components: Yield Curves, Loans, Assets and Technologies with variables for purchase cost, running cost, migration cost, service time, further detailed into interest rates, decrease in storage cost per storage unit, a “short-term-ism" factor, planning horizon
Type of tool	Simulation tool, based on Monte Carlo simulation, implementation unknown, maybe run on Prism
Availability of tools	None. A description of the model is available at: http://blog.dshr.org
References	Rosenthal, D.S., Rosenthal, D.C., Miller, E.L., Adams, I.F., Storer, M.W., Zadok, E., 2012. The economics of long-term digital storage, in: Memory of the World in the Digital Age Conference, Vancouver, BC. Retrieved from http://www.lockss.org/locksswp/wp-content/uploads/2012/09/unesco2012.pdf Rosenthal, D., Economic model of Storage, September 2011: http://blog.dshr.org/2011/09/modeling-economics-of-long-term-storage.html Rosenthal, D., Economic model of Storage, November 2011: http://blog.dshr.org/2011/11/progress-on-economic-model-of-storage.html

Table 12—EMLTS characteristics

4 Stakeholders Financial Information Requirements

In order to establish stakeholders' needs regarding financial information and cost & benefit modelling we analysed the results of an initial web based stakeholder consultation[15] conducted by the 4C project in conjunction with a 4C deliverable report [4C, D2.1, 2013] and also supported by partners in the 4C project, who have experience with identification of stakeholders' requirements from cost model development.

A brief description of the stakeholder consultation is shown below followed by a detailed analysis of each section of the consultation questionnaire and conclusions about identified needs. A list of the questions can be found in Appendix A.2.

4.1 Stakeholder consultation

The consultation took the form of a set of questions presented to a previously collated contacts list in order to establish their current practices relating to the assessment of digital curation costs and to obtain additional information regarding their most prominent challenges and needs in this work area.

Invitations for the web consultation were sent out to 296 stakeholder contacts. There were a total of 164 responses (55%).

The consultation had three sections. The first two sections consisted of questions relating to general information about the contacts organisation and information about the contacts themselves. The third optional part consisted of additional questions about stakeholders needs for financial information and current cost modelling practices. The two first sections were completed by 76 participants (46% of the responses). The full questionnaire including the third section was completed by 46 participants (28% of the responses).

The questions of the consultation were divided into the following sections:

1. Organisation (general information)
2. 4C participation (basic questions about sharing cost information, contact information)
3. Digital curation (Activities, Content, Accounting & budgeting, Cost modelling)

The first section covered basic information about the organisation with questions about core business activity, users, main funding sources and number of employees. Participants were asked to self-select the stakeholder group they identified themselves with (see Table 13 below). The next section dealt with questions about participation in the 4C project, sharing cost information and about contact data for further activities. The third section included information about digital curation and was further divided into subsections. To identify the stakeholder's needs for financial information the questions from this third part have been analysed in detail.

The questions had pre-defined answers to choose from, either single or multiple choice, but most also gave the option for providing comments. Some of the questions and answers allowed us to directly extract stakeholders' needs, whereas others required some interpretation. For example, the question "For what purposes does your organisation need financial information related to digital curation?" (Q28) and the answer "Budgeting", reveals a need that can be directly transferred into requirements for models. However, the question "Select the 3 main reasons for your organisation to use a cost model" (Q34) and the answer "To inform decision-makers" only indirectly expressed a need.

[15] 4C stakeholder consultation, http://consult.4cproject.eu/index.php/949288/lang-en [Accessed 31 Jan 2014]

The complete list of stakeholder's needs identified in each subsection can be found in Appendix A.3—List of Stakeholders' Needs.

What is the description that best fits your organisation?		
	Count	Percentage
Research funder	4	5·3%
Big data science	5	6·6%
Digital preservation vendor	7	9·2%
Government agency	10	13·1%
Publisher or content producer	3	3·9%
Data intensive industry	5	6·6%
Memory institution or content holder	18	23·7%
Small or medium enterprise	2	2·6%
University	11	14·5%
Other	11	14·5%

Table 13—Stakeholder groups represented in the consultation

4.2 Analysis of stakeholders' needs

The subsections used in the analysis were:

- Activities
- Content
- Accounting and budgeting
- Cost modelling

4.2.1 Activities

The questions in this subsection covered funding sources, annual budget, importance of digital curation activities within the organisation, out-sourced activities, infrastructure for digital curation, number of accesses and current breakdown of the activities.

The question addressing the importance of the activities ("How important are the digital curation activities when compared with your other business activities?") was answered by 61% of the participants that it is "a core activity". The complete results can be seen in Figure 3. Hence, for the majority of the respondents in this section digital curation is assumed to be the main activity within their organisation.

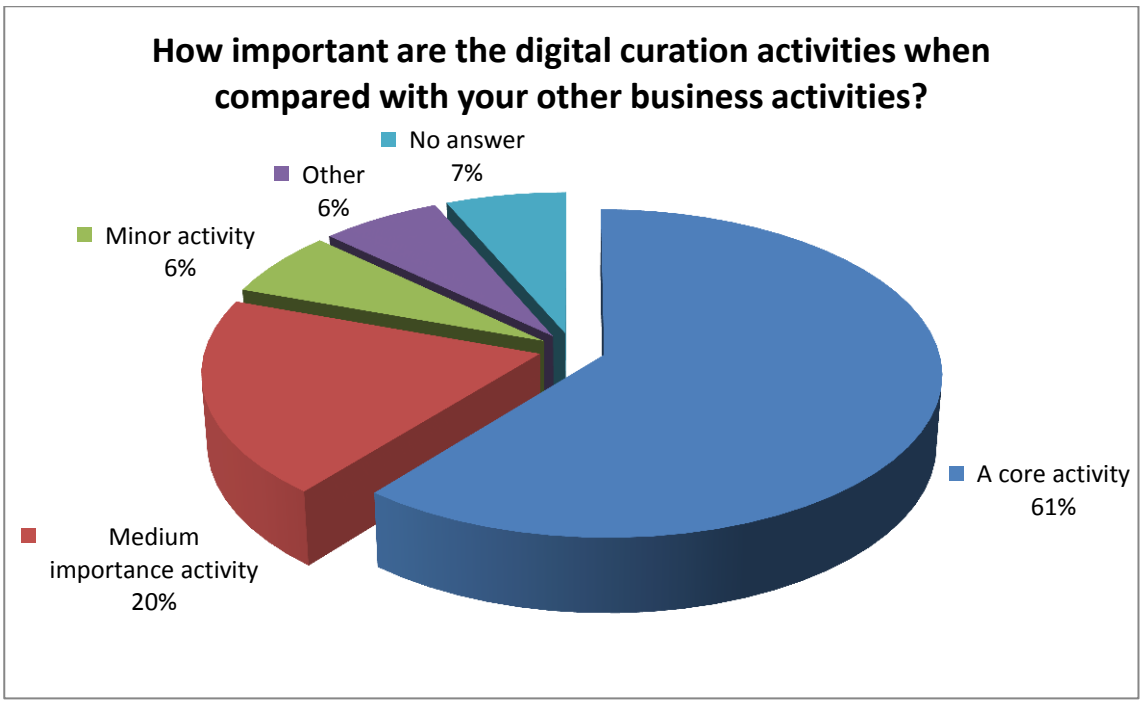

Figure 3—Importance of digital curation activities to stakeholders

Most of the organisations performed their digital curation activities in-house but some also outsourced parts of these activities. Of the 46 participants that completed the whole questionnaire 61% of them stated that activities were performed in-house, 24% partially and 7% completely outsourced them as shown in Figure 4. Different activities were (partially) outsourced, for example, digitisation or hosting of the repository.

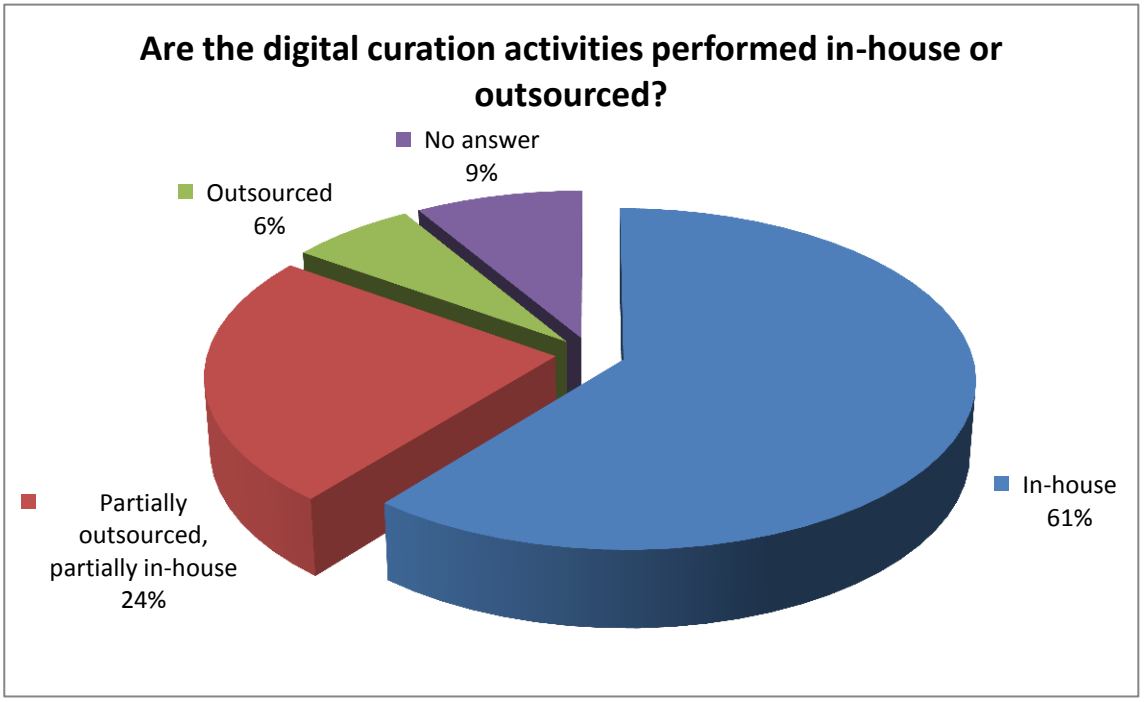

Figure 4—Digital curation activities, in house or outsourced?

The question regarding the breakdown of the costs revealed that from 46 answers 48% stated that the cost of their digital curation activities were not separated from other business activities and 15% stated that their activities were separated from other business activities but not further broken down (see Figure 5).

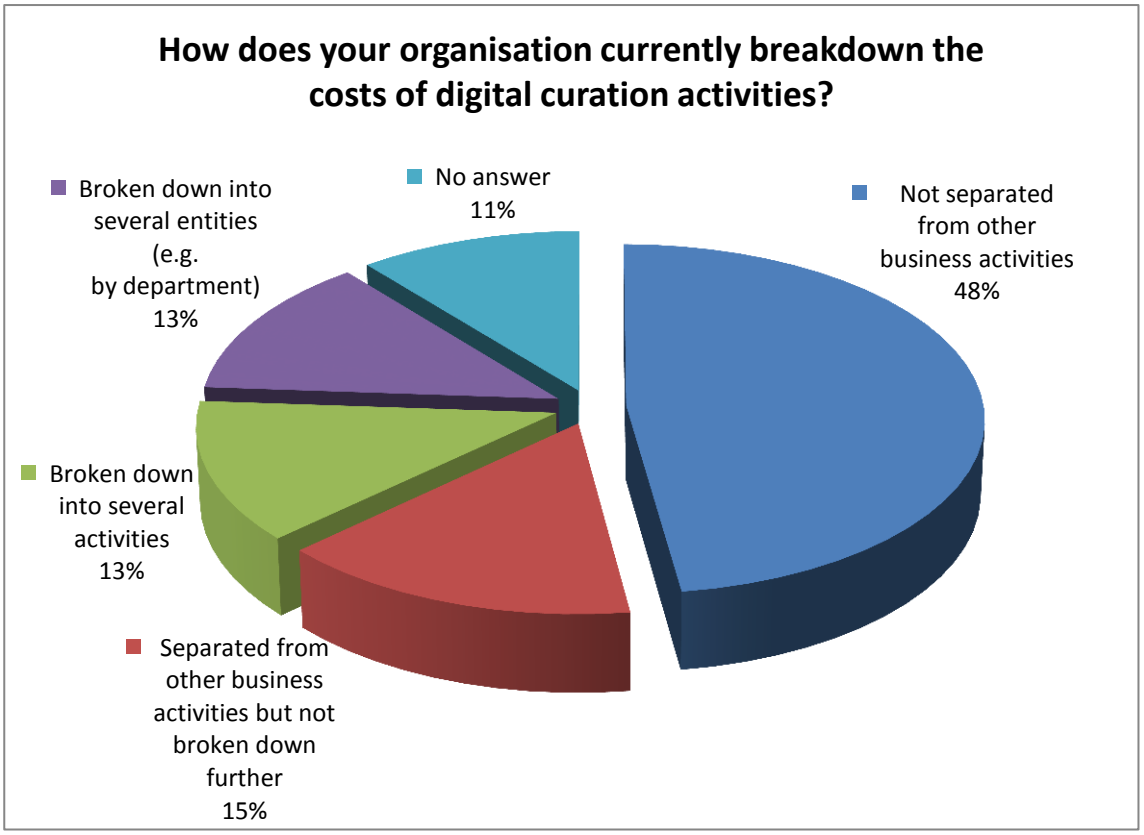

Figure 5—Stakeholders current practice in breaking down costs

Although most of the organisations are apparently aware of the importance of digital curation there is no consistent breakdown of its activities and its costs. The costs are often not distinguished separately and are encompassed in other business units.

4.2.2 Content

In this subsection further information about the digital assets of the participants of the consultation was gathered. It consisted of questions regarding content type, motivation for keeping assets, the benefits they represent, the timescale of preservation and the volume of assets. The participants were asked about the type of information assets and multiple response options could be chosen. The analysis of the answers shows that the content is diverse and range from records of business activities through cultural heritage data to research data and scholarly publications. In most cases the contents consists of digitised materials (see Figure 6).

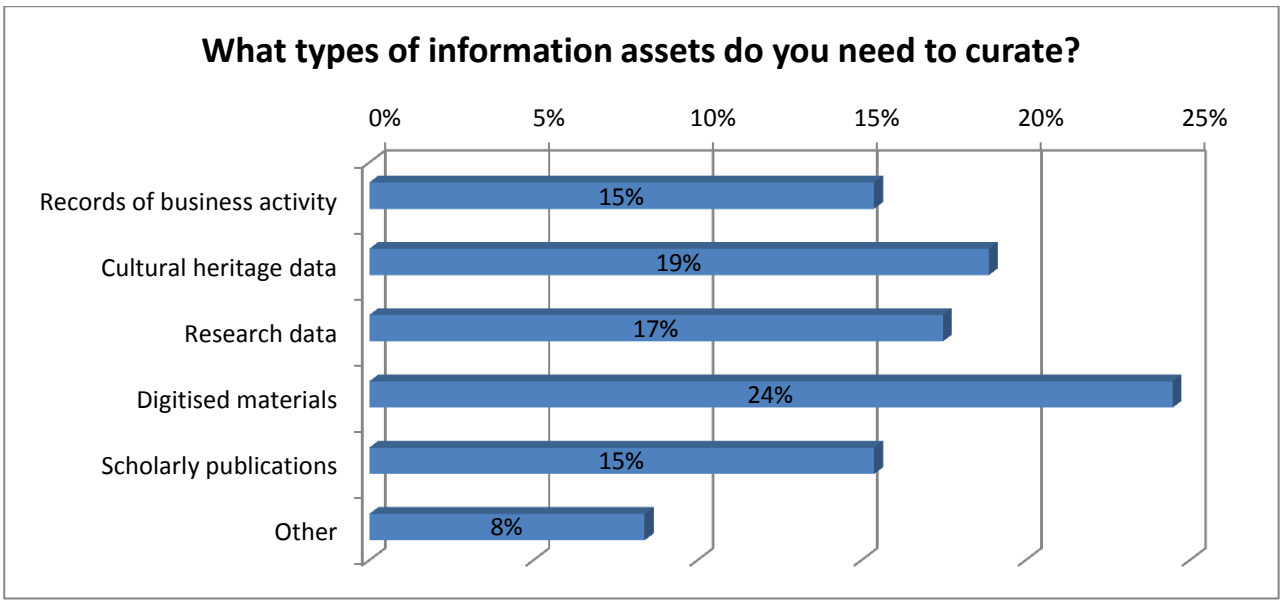

Figure 6— Distribution of given answers about the type of curated assets.

This is a vital piece of information to know if a given cost model can handle the required type of assets.

The question "What is the motivation for keeping these assets" could also be answered with multiple responses. A total of 87 answers were given with the following distribution:

- ensure availability of public good (34%)
- legal requirement (32%)
- business requirement (23%), and
- other (10%)

Some differences between the stakeholder groups could be found in the motivations for keeping digital assets. "Publishers or content providers" and the "Data intensive industry" were for example driven by business and legal requirements whereas other groups were mainly motivated by ensuring the availability of the assets for the "public good".

With regards to the benefits that the assets represent to an organisation most participants, 37 out of the 46 that completed the questionnaire, selected "Fulfil the institutional mission". The stakeholder groups of "Data intensive industry" and "Publisher or content provider" saw additional benefits in providing monetary profit or reducing costs in the long-term.

When questioned about the timescale of curation 70% stated that they had to maintain access to the assets selected for a "long term storage (infinite)" period. Only 13% needed to maintain access over a "medium term storage (5-20 years)" period (see Figure 7).

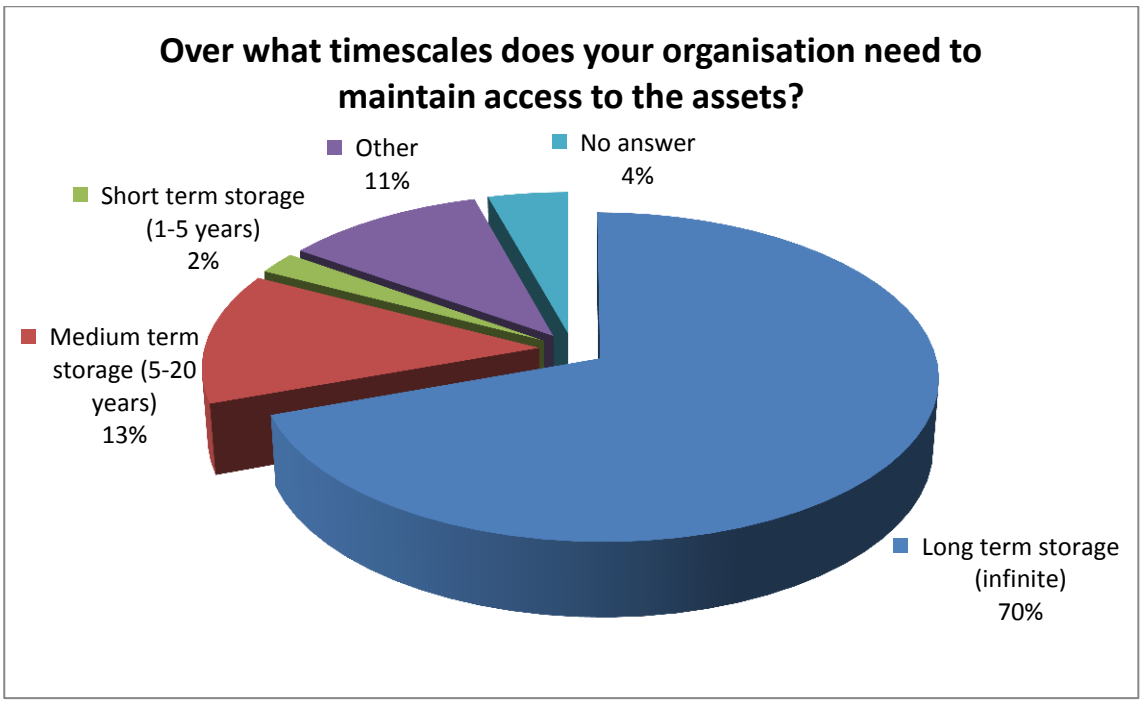

Figure 7—Timescale to maintain access to assets

The following questions dealt with the number and volume of files with the interviewees being asked to estimate current volumes and the expected increases in 5 years. The answers showed a diverse result with no bias towards the quantity of files (see Figure 8 and Figure 9).

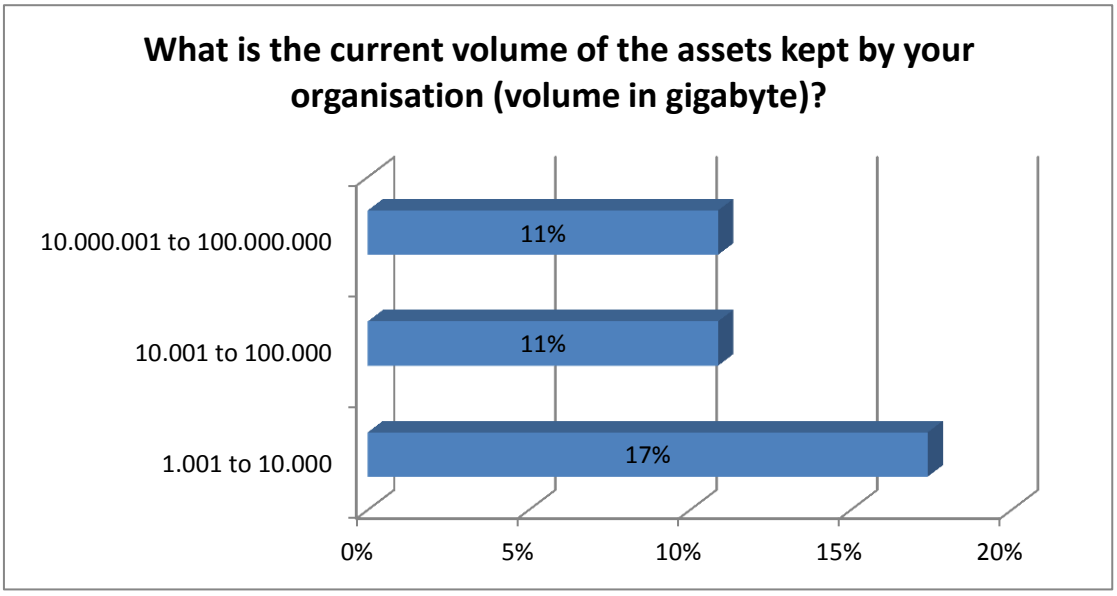

Figure 8—Volume of assets kept by organisations showing the 3 most frequent responses

The need to ensure sustainability of digital assets emerges from the long term storage requirement. Furthermore information about the relationship between the quantity of assets and the cost is needed by the stakeholders because the number and size of files of their assets are diverse.

The stakeholder consultation showed that cost models should be able to handle “required types of assets” as an input, and that the timescale for which digital assets need to be curated had a focus on long time periods. Also it showed that the number of files and their volume are diverse, and as a consequence a potential cost model need to be scalable from small collections to large archives.

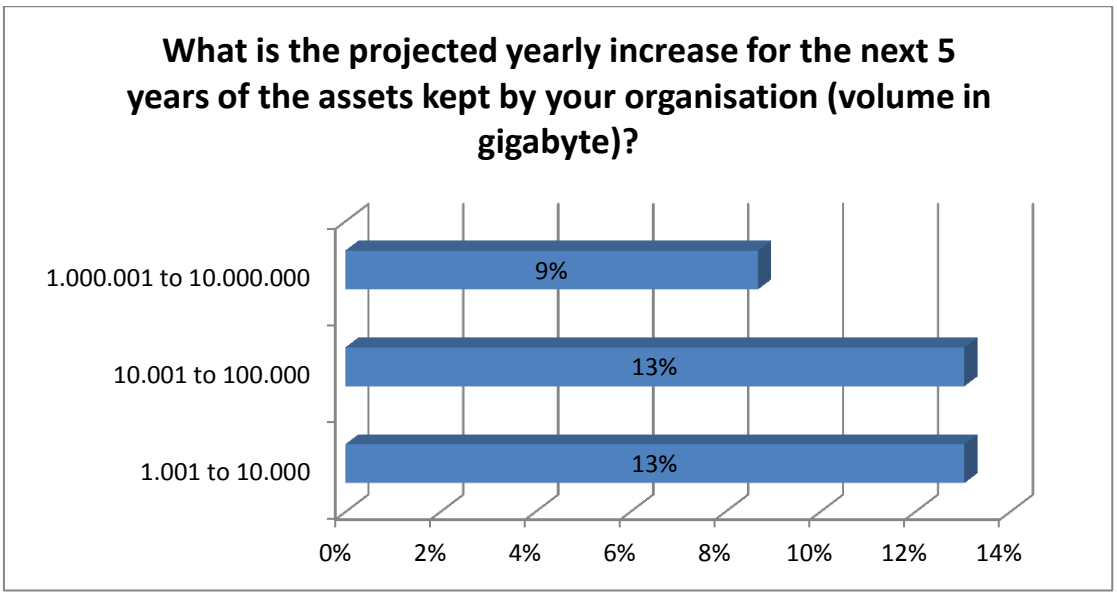

Figure 9—Projected volume increase for the next 5 years showing the 3 most frequent responses

4.2.3 Accounting & budgeting

In this subsection the stakeholders answered questions about the reasons why financial information relating to digital curation was needed. The reasons selected most by the 46 participants were associated with the need for budgeting and comparing cost and benefits of alternative scenarios to support decision-making (see Figure 10).

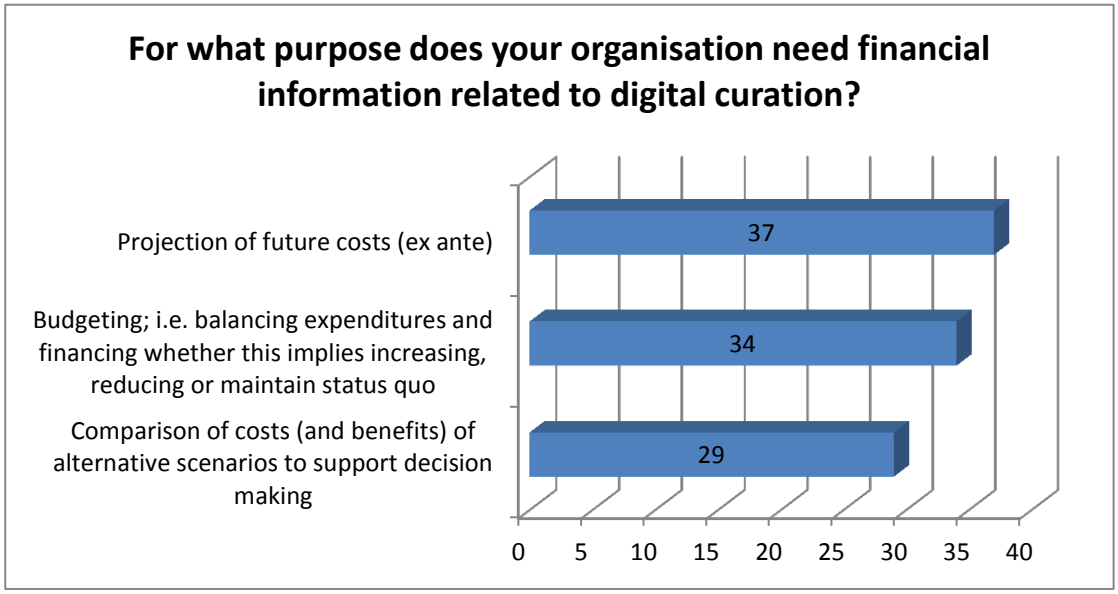

Figure 10—Need for financial information

The consultation also showed that in 28 organisations the department director was responsible for accounting and budgeting for digital curation. In contrast, in 16 cases it was the repository manager and the general financial/accounts manager's task. Thus, different target audiences with different levels of knowledge about digital curation were responsible for the task of accounting and budgeting for digital curation within the organisations, and this calls for models that facilitate communication between these groups and emphasises the importance of a clear documentation of the models.

Further needs identified in this part of the questionnaire were:

- adaptable tools
- clear definitions of activities, and
- guidance/best practice approach.

The question “What type of costs does your organisation need to account for” was answered differently by the groups of stakeholders. “Full economic costs, total costs of ownership, lifecycle costs” was chosen by all representatives of the group of “Data intensive industry”. In contrast, the group of “Publisher or content provider” appeared to have no interest in this type of costs. Similarly, the “Government agency” group were concerned with “Investment costs”, whereas the “Data intensive industry” group showed no interest in these types of costs. These differences between the interests of the stakeholder groups show that there is a need for cost models that are flexible enough to take into account these different interests in the type of costs that need to be accounted for.

4.2.4 Cost modelling

In this part of the consultation the participants were asked to select the three main reasons for using a cost model (see Table 14). The main motivation selected by 36 out of 46 participants was “To inform decision makers” which implies a need to account for the benefits and value of alternative scenarios. The second most popular answer selected 33 times was “To find out costs of preserving assets”, and the third reason selected by 28 was “To ensure the efficient use of resources”, again an expression of the need to evaluate alternatives.

Select the 3 main reasons for your organisation to use a cost model	
To inform decision makers	36
To find out the costs of preserving assets	33
To ensure the efficient use of resources	28

Table 14—Main reasons for stakeholders to use a cost model

Likewise, the participants were asked to indicate which statements matched their reasons for selecting and using a cost model (see Table 15). The three most popular selections were “Is the model easy to use and adaptable”, selected 30 times, followed by “Model has been validated by a similar organisation in your sector” selected 29 times and “The scope of the model” selected 26 times. From these answers we derived:

- the need to trust the model and know the principles it builds on
- the need to ensure sustainability of digital assets
- the need for enhancing the efficiency and ease of use of the model, and
- the need to obtain cost figures in an automatic and consistent way.

On what basis would you select a cost model?	
Is the model easy to use and adaptable	30
Model has been validated by similar organisation in your sector	29
The scope of the model; e.g. covering the digital curation lifecycle	26

Table 15—Reasons to select a cost model

The final set of questions covered former experiences with cost models, possible improvements, origins of the models and the granularity of activity grouping. Answers were consistent with the results from the previous questions and included difficulties with not easily adaptable models, insufficient definitions of

activities, no assessment of the quality of activities, lack of usability and missing guidance & indications of best practice approaches.

Answers in this subsection showed no distinct differences among the stakeholder groups which implies that the questions asked address fundamental principles of cost modelling which concern all stakeholder groups.

5 Gap Analysis—Models' capabilities versus stakeholders' needs

The goal of the gap analysis was to identify areas in current cost & benefit modelling that, from a stakeholder's perspective, should be improved to support the uptake and use of models. Further the goal of the analysis was to identify features of the models that might be appealing to users and provide good practices for model developers.

5.1 Method

To identify gaps in the existing models we evaluated the models capabilities against various stakeholders needs. The models capabilities were evaluated based on the provided descriptions of the models (see Section 3 'Description of Existing Models'), available model documentation and through the use of the tools themselves. The stakeholders' needs were derived from the stakeholder consultation (Section 4 'Stakeholders Financial Information Requirements') as well as from model developers within the 4C consortium who had experience with specifying requirements for cost models.

In this context we define a gap as a shortcoming between users' needs and the capabilities of the models. As such, the possible types of gaps we identified include:

- Gaps in individual models, for example "model 'X' does not specify indirect costs"
- Gaps in the collective mass of models, such as "Only very few models are both cost and benefit models"

Gaps in individual models are particularly interesting for models users, whereas gaps across the full set of models as a whole are more interesting for model developers. For the purpose of this report, if less than half of the evaluated models do not handle a requirement this is considered to be a gap.

The gap analysis was coloured by the fact that the models differ considerably in scope and design. Furthermore, the stakeholders have different needs depending on the context they operate in. For these reasons we didn't set out to rate the models' effectiveness, but rather to enable a comparison of specific characteristics of the costs models' functionality and to identify gaps that will require further investigation before increased uptake amongst stakeholders is realised.

While the stakeholder consultation provided insights into what model users expect from using cost models, some of the needs identified through the consultation were expressed in a rather abstract form for example, "To inform decision makers" or "To find out the costs of preserving assets". To enable an evaluation of models' capabilities, these abstract needs had to be transformed into more concrete and measurable requirements for models. This transformation brought out the underlying assumptions and preconditions of the needs. For example, the need to assess the cost of preserving objects requires that the model covers a range of curation activities, handles the requested types of assets and the envisioned preservation strategies, and includes some means of estimating future costs.

5.1.1 Model evaluation schema

In order to evaluate the models' capabilities we designed a schema in which stakeholders' needs were transformed to requirements for models and expressed as Boolean (yes/no) questions. Altogether we identified 79 requirements (ID1-79).

Based on the initial investigation and summary of models we grouped the identified requirements in the schema under four key model characteristics (see also Section 2 'Characteristics of Cost and Benefit Models'):

1. Model type
2. Cost structure (activity and resource)
3. Cost variables
4. Usability

The individual requirements are described in more detail below.

The model evaluation was carried out by two partners of the 4C team to ensure that there was consistency in our approach—one partner did the evaluation and another the review of the results. The evaluation results were also sent to the model owners so that they could validate or challenge any of our findings. The analysis of gaps and provision of recommendations was then done collaboratively by the 4C team.

5.1.1.1 Model type

Requirements under this characteristic determined whether each model was an economic model (ID1), a benefit model (ID2-4), or a cost model (ID5). If the model was a cost model it was also evaluated to see if it was intended for accounting (ID7) and/or budgeting (ID8), and if that was the case, if it included a means of estimating costs over the short, medium or long term (ID9-11). These were the only characteristic that dealt with benefit models and economic models, the rest of the schema was only relevant to cost models.

5.1.1.2 Cost structure (activity and resource)

This characteristic included the requirements for breaking down the costs by activity and resource. It evaluated which resource types are accounted for, including capital/investment costs, maintenance and operation costs, indirect costs and labour costs (ID12-16). It also evaluated if the applied structure was based on a standard or if it was customised (ID19-20). It further evaluated whether the models break down costs by activity, organisational unit, or other functional criteria (ID22-24). Since many models build on the OAIS Reference Model and in order to capture the level of detail of the models, it also evaluated which of the OAIS entities the models accounted for (Ingest, Data Management, Archival Storage, Access, Preservation Planning, Administration, Common Services) and whether any of these OAIS entities were broken down even further (ID22-43). To cover all digital curation activities it also evaluated if the models accounted for pre- and post-repository activities as well as management.

5.1.1.3 Cost variables

This characteristic covered requirements regarding how cost variables are modelled. It evaluated several service adjustments, related to the assets and the curation systems and services. It evaluated if the models could divide the costs by amount (ID21), if quantities could be expressed as number of items or by volume (ID 44-45), if it was possible to indicate estimated yearly increase in quantity (ID46-47), and if the models assumed a minimum or maximum amount of assets for the cost assessment (ID53-54). In addition this part of the schema evaluated if the models could handle one or more simple data formats with few levels and dimensions, for example 2D data (documents, images, sound, video and so on.), and if they could handle one or more complex formats with many levels and dimensions, for example chemical or meteorological models (ID49-52). It also evaluated if the models could handle various aspects of the quality of the curation systems and services, such as specification of requirements for upload/download

amount and frequency (ID48) as well as if the models could handle different preservation strategies, such as migration at different points of the lifecycle (normalisation or on demand) and emulation (ID55-58), and if they enabled specification of the quality of activities or of the repository system as a whole (ID59-61). Finally this characteristic also evaluated if the models included means of economic adjustments, such as depreciation and discounting (ID17-18).

5.1.1.4 Usability

This characteristic included requirements regarding the functionalities and user-friendliness of the models. It assessed if the model was implemented as an electronic tool, and if so which type, spreadsheet, web application, or another application, and if the tool has a graphical user interface (ID70-73). This was followed by evaluating if the models included pre-defined entries, such as activity checklists (ID6), parameters and values, and if these could be changed to cater for specific scenarios (ID77-79). Along this line it evaluate the efficiency of the models expressed as the time it takes to get results of a calculation, ranging from 1 hour to half a day (ID68-69). It also evaluated the models design, if it was modular and easy to add or remove components and if it applied algebra formulas to support calculations (ID74-75).

An additional evaluation was made regarding the quality of the documentation associated with the model, both that relating to model design (ID76) and user-guides, such as manuals, presentations, papers (ID64). Finally there was also an assessment of the intended users of the models, are they managers of systems and services or account managers (ID62-63) and it evaluated the models learning curve, the time it would take users familiar with the curation domain to learn to use the model, ranging from 1 day to 2 weeks (ID65-67).

5.2 Results

First, for each model characteristic (Model type, Cost structure, Cost variables and Usability) we present the detailed results of the evaluation of the models against stakeholders' needs and the identified gaps. In addition a series of incentives—expressed as drivers—for the uptake of models are provided. A condensed version of the full individual result reports from the gap analysis can be found in Appendix A.4.

We then present a use case test to assess how useful the current models might be for UK Universities wishing to assess in-project and post-project curation costs associated with new research grant applications.

The identified gaps and drivers inform the discussion in Section 6 'Discussion and Recommendations of this report', and are synthesized in Section 7 'Conclusions'.

5.2.1 Detailed gap analysis

5.2.1.1 Model type

The characteristic "Model type" addresses three types of models: The economic model, the benefit model and the cost model, described in Section 2 'Characteristics of Cost and Benefit Models'.

Identified gap
Lack of benefit models

All the evaluated models are cost models, two of which have benefit modules incorporated within them[16]. When matching the stakeholders' needs against the model capabilities, we identified one important gap, namely the lack of benefit models. The evaluation of each requirement is shown in Table 16 and described further below.

	Model									
Model type	T-CMDP	NASA-CET	LIFE3	KRDS	CMDA	CMDP	DP4lib	PP-CMDS	CDL-TPC	EMLTS
Economic model	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘
Cost model (CM)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CM, with benefit module	✘	✘	✘	✓	✓	✘	✘	✘	✘	✘
CM, with activity checklist	✓	✓	✓	✓	✓	✓	✓	✘	✘	✘
CM, past/current costs	✓	✓	✓	✓	✓	✓	✓	✘	✘	✘
CM, future costs (fc)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CM, fc—1-5 years	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CM, fc—5-20 years	✓	✓	✓	✓	✘	✓	✘	✓	✓	✓
CM, fc—more than 20 years	✓	✓	✓	✓	✘	✘	✘	✓	✓	✓

Table 16—Results of the model evaluation for the characteristic "Model type"

Cost model

The stakeholder consultation showed that the majority of the participants envisioned that their organisation would be likely to profit from digital curation cost modelling. However, only 15% of the participants indicated that they had tried to use a cost model.

Cost model with benefit module

Only two of the cost models reviewed account for benefits and value. The KRDS Benefits Framework Tool identifies benefits and the KRDS Value-chain and Benefits Impact Tool help identify potential measures or illustrations of the value and impact of those benefits. These two tools are meant to be used in conjunction with the KRDS activity based cost model. The CMDA includes a balanced scorecard approach to ensure that the mission of an organisation and existing strategies are translated into strategic objectives that can be measured operationally.

The stakeholder consultation revealed a clear demand for assessing benefits, and consequently there is a gap between most of the existing models' capabilities and the stakeholders' needs. Stakeholders indicated that they need a model for comparison of costs and benefits of alternative scenarios to support decision making (67%) and affirmed that they would use the models if these were capable of assessing the benefits and the values of digital preservation (63%).

A lot of the consulted organisations do not obtain any monetary benefits by curating their assets, and a model displaying their intangible benefits (for example the value of documentation, the value of fulfilling a

[16] The reason for noting if the model is an economic model was that the Economic Sustainability Reference Model (ESRM), which is an economic model, was initially included in the evaluation.

mission) would help in understanding the raison d'être of these organisations, in turn facilitating funding and increasing reputation.

Organisations that are looking to obtain monetary profit logically seek a model to help them specify the profit. This is also true for organizations that are evaluating the possibilities of outsourcing activities and for organizations that seek to put a price on paid services.

Driver 1:

The ability of cost models to assess associated benefits supports the need for justifying and sustaining costs, and enables comparing costs and benefits of alternative scenarios to support decision-making.

Cost model with activity checklist

The majority of models reviewed do provide an activity checklist to help users break down costs across the curation lifecycle, and many of these are based on the OAIS functional entities. CMDA includes the DANS-Activity-based Reference Model (DANS–ABRM) which describes activities taking place in a trusted repository. Other models provide detailed activity checklists for specific sections of the curation lifecycle. For example, CMDP provides very detailed activities around archival storage. Some models provide an opportunity for users to define their own checklist of activities. For instance, the CDL-TCP model provides some defined activities, however these are not presented as a checklist, but as the so-called 'Intervention' sheet that allows the user to enter activities themselves.

A little over half of the of the consulted stakeholders (53%) indicate a need for a checklist of asset management activities that incur cost in order to identify which costs are included and which are not.

Driver 2:

The availability of activity checklists in cost models supports the need for guidance on specification and definition of digital curation activities that incur cost.

Cost model—past and current costs

Most of the tools offering pre-defined data and formulas can handle past costs but are designed as planning tools that look ahead rather than back. Some tools are better able to record past costs such as LIFE3 which allows users to record procurement costs as part of ingest. The models that do not provide pre-defined data and formulas can be applied to any activity whether current, past or future (KRDS for example).

When asked "For what purposes does your organisation need financial information related to digital curation?" about 1/3 of the stakeholders stated that they need it for accounting (calculation of past costs). Furthermore, some stakeholders indicated that they need financial information for internal financial management (43%) and for external legal requirements (16%).

Driver 3:

The ability of cost models to account for past costs (ex-ante) supports the need to meet internal management requirements as well as external legal requirements.

Cost model—future costs

All of the models will help you to plan for short term future costs. As noted above, some models provide pre-defined data and formulas while others expect users to develop these themselves. Predicting future costs over the midterm and longer can be difficult with pre-defined data and formulas, as some of the

models do not reflect likely changes such as annual pay increases or changes in staffing. For example, T-CMDP allows you to include a 'repeat after M years' calculation. However, this assumes that the same activities will be carried out by the same roles so doesn't really allow for changes in staffing and/or pay. Several tools can handle costs for the longer term, but many also state that forecasts aren't as accurate beyond a certain time, for example LIFE3 indicates ten years from the point of ingest. The EMLTS model estimates how storage costs might vary over a 100 year period and aims to reflect the effects of changing technology over long periods of time.

Users need to know the cost of curating and sustaining access to digital assets over time. Thus, the most popular (80%) reason that respondents selected for using cost models was for budgeting (projection of future costs).

Driver 4:

The ability of cost models to estimate future costs (ex-ante) supports the need to make budgets and balance potential costs and revenues.

5.2.1.2 Cost structure—Resource

This characteristic evaluates how the models breakdown costs by resource type and whether the models apply depreciation/amortization and discounting (see Section 2 'Characteristics of Cost and Benefit Models' for definitions).

Identified gap
(none)

When investigating the stakeholders' needs in relation to the capabilities of the models to breakdown costs by resource (see Table 17 below), we found that there are no significant gaps. However, the models do not break down costs of resources in the same way, which makes it difficult to compare cost data extracted from different models.

	Model									
Resource breakdown	T-CMDP	NASA-CET	LIFE3	KRDS	CMDA	CMDP	DP4lib	PP-CMDS	CDL-TCP	EMLTS
Capital (investment) costs	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓
Maintenance/operating costs	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓
Indirect costs	✓	✓	✓	✓	✓	✗	✓	✗	✓	✗
Labour costs	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗
Differentiate labour costs	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗
Depreciation/amortization	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓
Discounting	✗	✓	✓	✓	✗	✗	✗	✗	✓	✓

Table 17—Results of the model evaluation for the characteristic "Cost structure– Resource"

Direct costs (capital, maintenance and operation costs)

Most of the models reviewed do allow users to account for capital (investment), maintenance and operating costs. The only exception is the PP-CMDS, which provides the total cost based on experience from storage providers and a few large-scale institutions.

In general stakeholders indicated a stronger need for accounting for running costs than investment costs. 65% of the stakeholders stated a need to account for maintenance and operation costs, and 35% for capital (investment) costs.

Indirect costs

With some of the models, we did not have access to cost data or tools so it was difficult to assess whether indirect costs could be handled. For those that do cover indirect costs, the majority included aspects relating to office space for staff and equipment. T-CMDP Includes 1.25 x salary to provide full cost of labour and also include 20% overheads for office space calculations. Other models made use of Full Economic Costing (FEC) for calculating indirect costs. LIFE3's default setting is not to capture FEC, but users can opt to calculate costs based on FEC. KRDS uses the Transparent Approach to Costing (TRAC)[17] approach to FEC. CMDA aims to record as many activities as 'direct' costs as possible and a cause-and-effect criterion was applied to identify the cost-allocation base for each indirect-cost pool.

50% of stakeholders indicated a need to account for indirect costs.

Labour costs

Most of the models allow users to differentiate between different types of labour costs. However, in some models there are assumptions made about which staff may be involved. For example, T-CMDP assumes that there is a dedicated Archive and Preservation facility with a skilled team, familiar with digital preservation, aware of IT and record keeping issues'. The specific roles used in the model include administrative assistants, record keepers, supervisors, data entry operators, software engineers, and senior ICT person. In other models, labour costs are defined within a project rather than an institutional framework (LIFE3 being a case in point) which uses default roles in including Senior Manager, Project Manager, Technical Officer, Project Team, and Operational staff. The LIFE3 model allows users to choose to assess costs by year, day/hour/minute rates to calculate costs for shorter activities. In the CMDA, staff are characterised as either ICTa or ICTb. ICTa represents the more permanent, lower payment rate employees while ICTb contains mainly outsourced experts with a much higher payment rate. The CMDP characterises staff costs by salary level rather than role (low, medium, high salary) and allows these levels to be changed according to local needs.

In the consultation stakeholders were not asked about their need to express labour costs. However, 46% of the stakeholders stated a need to record full economic costs (FEC) and thus indirectly also to account for labour costs.

Financial adjustments

The KRDS guidance explains each type of financial adjustment and users of the cost model will implement these as a spreadsheet, populated with data and adjustments agreed by the institution. Most of the models allow for depreciation of capital costs. T-CMDP assumes a depreciation of 33% per year for hardware and software. The PP-CMDS model allows the user to define how costs per unit change over time. Some other models allow users to define their own depreciation variables. Some of the models also allow for discounting. LIFE3 includes a cost deflator variable that can be specified in the 'model variables' tab.

The consultation did not survey the need for financial adjustments. However, one respondent suggested the inclusion of "economic adjustment" to improve the cost models.

[17] http://www.jcpsg.ac.uk/guidance/

Driver 5:

The ability of cost models to breakdown curation costs by resource in a standardised way supports the need for comparing costs of alternatives.

5.2.1.3 Cost structure—Activity

This characteristic covers stakeholders' needs regarding the way the model breaks down costs by activity and at what level of detail the costs are broken down. It also evaluates if the cost breakdown is based on a standard and if it allows for customisation or not (see Table 18).

Identified gap
(none)

We did not identify any important gaps in the models' ability to breakdown costs by activity. However, as seen with the breakdown by resource, there is no consensus as to how the models breakdown activities, which makes it difficult to compare the output of different models.

	Model									
Activity breakdown	T-CMDP	NASA-CET	LIFE3	KRDS	CMDA	CMDP	DP4lib	PP-CMDS	CDL-TCP	EMLTS
Standardised	✓	✓	✓	✓	✓	✓	✓	✘	✓	✘
Customisable	✓	✓	✓	✓	✓	✓	✘	✓	✓	✓
By activity	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
By OAIS entities	✓	✓	✓	✓	✓	✓	✓	✘	✓	✘
By OAIS sub entities	✓	✓	✓	✓	✓	✓	✓	✘	✓	✘

Table 18—Results of the model evaluation for the characteristic "Cost structure– Activity"

Standardised structure

Standard approaches for describing the lifecycle of digital curation, the range of individuals involved in creating, managing and preserving access to digital assets, and curation use cases are needed to allow model users to be able to interpret results obtained from different models and to carry out some comparison between scenarios.

Most of the models we reviewed do make use of a widely used or standardised activity breakdown structure. In many cases, models made use of OAIS Reference Model and included a defined activity model. Perhaps one of the most defined activity models we encountered was that of NASA-CET. With the NASA-CET activity model, users must map planned activities to the CET Data Service Producer (DSP) reference model. In the case of CDL-TCP some OAIS terms have been renamed to facilitate understanding by non-specialists.

Most stakeholders stated that the cost models they had tested did not cover all the activities that they needed to cost digital curation in their organisation and also not at the right level of detail.

Customisable structure

Several of the tools encourage end users to modify the model structure to reflect local circumstances (for example LIFE3). With KRDS, users are encouraged to adapt the language used in the model and benefits

spreadsheets to reflect local strategies and objectives. Steps 4, 5 and 6 in their user guide deal with adapting model for local use.

Model users want to be able to adapt cost models to reflect their particular organisational environment. However, any customisations also represent a trade-off regarding the model's ability to output data that is comparable across organisations or scenarios.

Breakdown by activity

All models divide costs by activity, however at different levels of detail. In some cases organisational entities, such as departments, represent an activity.

Many current cost models only calculate the cost of curation from the point of ingest into an archive. However, with increasing mandates to retain research data, many research intensive Universities must now make informed decisions at the grant application stage about both the in-project data management costs and their potential to retain the data in house for as long as need be, in most cases at least ten years. This will involve being able to identify directly incurred costs during the active phase of the research project as well as assessing the potential budgetary impact on retaining the data over the longer term. Institutions will want light-touch options as well as resource-intensive options to cover the vast array of data they will need to preserve access to.

Breakdown by OAIS entities and sub entities

As stated above most of the models use the OAIS Reference Model as the basis for breaking down costs. The OAIS entities include Ingest, Data Management, Archival Storage, Preservation Planning, Access, Administration, and Common Services. Beyond these repository entities, digital curation includes pre-repository (production, pre-ingest) and post-repository (use and reuse) activities, as well as general management. Most of the models cover these entities; Archival Storage and Ingest are especially well represented whereas production activities, such as digitization costs, are less well covered. Many of the models also divide entities in sub activities (see Appendix A.4 for details).

We received a comment regarding the level of cost breakdown and accuracy from Stephen Abrams (CDL-TCP model owner): "We purposefully did not attempt to model costs at a finer degree of granularity, such as would be required to break things down at the sub-OAIS entity level. We believe, perhaps somewhat paradoxically, that past a certain level of modeling granularity the accuracy in estimating costs actually decreases as the granularity increases. (In essence, we feel that it is easier to make an accurate estimate of time in terms of days rather than hours, weeks rather than days, etc.) We have tried very hard to ensure that the TCP does not give the impression of greater accuracy than may be justified given the many assumptions and intuitive estimates that go into it. Also, we found in many cases that it was difficult to map our local practices into the OAIS sub-functions in an obvious and unambiguous manner." [18]

Driver 6:

The ability of cost models to breakdown curation costs by activity in a standardised way supports the need for comparing costs of alternatives.

[18] Email communication with Stephen Abrams, Associate Director, UC Curation Center, California Digital Library

5.2.1.4 Cost variables

In this category we analyse the models' ability to cover variables, which have an impact on the costs with a focus on the quantity and quality relating to information assets. The cost variables include number/volume of assets, types (simple/complex) of assets, quality of specific activities, repository system upload/download capacity, and repository quality in general. We also evaluate if the models have lower or upper boundaries of the amount of assets they can handle.

Identified gaps

A lack of models which address upload/download capacity

A lack of support for specifying the quality of repositories

The overview of the evaluation results is presented in Table 19. When compared with the stakeholders' needs we identified a lack of models that address the upload/download capacity of repository systems, and more importantly a lack of models that allow users to specify the quality of repositories.

	Model									
Cost variables	T-CMDP	NASA-CET	LIFE3	KRDS	CMDA	CMDP	DP4lib	PP-CMDS	CDL-TCP	EMLTS
Quantity of assets	✓	✓	✓	✓	✓	✓	✓	✓	✓	✘
Min/max amount of assets	✓	✘	✘	✓	✓	✘	✓	✓	✘	✓
Different types of assets	✓	✓	✓	✓	✓	✓	✓	✓	✓	✘
Quality of activities	✓	✓	✓	✓	✓	✓	✘	✘	✘	✘
Upload/download capacity	✘	✓	✓	✓	✘	✘	✘	✓	✘	✘
Quality of repository	✘	✘	✘	✘	✓	✘	✘	✘	✘	✘

Table 19—Results of the model evaluation for the characteristic "Cost variables"

Quantity of information assets

The overview of the quantity of assets is presented in Table 20 below. Seven of the models consider the number of assets in their calculations and eight models account the volume of assets. The CMDA model also allows the organisations to specify the number of privacy protected files. From the evaluated models the DP4lib and CDL-TCP models do not account for the number but rather the volume of assets. The EMLTS model focuses only on storage costs and does not refer to the number of assets. T-CMDP breaks down costs by amount of assets and calculates the batch costs by dividing staff costs across a number of items. The LIFE3 model has a 'refine creation' tab in which project variables are identified (quality and volume); in the 'refine bit-stream preservation' tab, costs are broken down by storage requirements in megabytes. KRDS incorporates this breakdown under 'Cost Drivers'.

	Model									
Quantity of assets	T-CMDP	NASA-CET	LIFE3	KRDS	CMDA	CMDP	DP4lib	PP-CMDS	CDL-TPC	EMLTS
Number of assets	✓	✓	✓	✓	✓	✓	✓	✓	✘	✘
Volume of assets	✓	✓	✓	✓	✘	✓	✓	✓	✓	✘
Yearly increase (number)	✓	✓	✓	✓	✓	✓	✘	✓	✘	✘
Yearly increase (volume)	✘	✓	✓	✓	✘	✓	✘	✓	✘	✘

Table 20—Results of the model evaluation for the characteristic Cost variables – Quantity of assets

Seven of the models allow for the estimation of annual increases in the number of assets. Five of the models cover the estimation of annual increases in data volumes. The NASA-CET model also addresses additional engineering effort required over a project. LIFE3 allows the organisations to specify increases in number and volume. KRDS covers increases in guidance but provides no formulas for calculations.

A yearly increase of assets is not supported by all evaluated models. The T-CMDP model supports defining costs incurred on a per-year basis as well as calculations over a timespan of multiple years. The assets are defined in number and size of batches per year. It is possible to specify the number of existing and new batches per year. The DP4lib model allows no specification of increasing asset sizes because it calculates its results only on a yearly base. The model lacks support for cost estimations for more than one year.

The CDL-TCP model does account for the volume of assets. Most routine preservation actions performed on content, such as characterization, fixity, normalization, and so on are supposed to be automated. As a consequence these costs are seen as independent from the number of assets they involve. Calculations using the model allow specification of time spans of several years but do not account for increases of the volume of assets.

The methods used by the models to account for the quantity of assets are different. They do this either by number or volume of the assets or both. If a model does not support yearly increases of the assets in number or volume (the DP4lib model for instance), organisations can incrementally calculate the results for each year and adjust the number or volume of the assets for each calculation step manually.

Maximum and minimum amount of assets

Only two of the models assume a minimum amount of assets. LIFE3 offers suggested volumes, which can be altered by the user, with the default setting for “low volume” being fewer than 100,000 items. CMDP similarly supplies default values that can be changed, but the storage costs are estimated based on systems with a capacity between 1-500 TB. The lowest volume specified in CDL-TCP is “up to 100GB” and KRDS makes no assumptions on the number of items.

The LIFE3 default setting for “high volume” is more than 1,000,000 items but the user can change this. A volume level of “up to 100TB” is the largest that can be specified in CDL-TCP. This restriction is not caused by the underlying model and is based on an arbitrary stopping point which has been chosen to keep the tool simple. Calculations can be extended easily above this boundary. There is no set limit in PP-CMDS , but the resources of the system on which the application is running may constrain the volume that can be modelled.

The stakeholders expressed the need to account for various collection sizes. If models have boundaries in the amount of data they can handle, this represents a limitation in their usability for the stakeholders.

Existing limits in the models are, however, rare and not part of the models themselves but rather part of the associated tools in order to keep them simple or to save resources of the computing system.

Driver 7:

The ability of cost models to estimate how costs scale with the quantity of information assets supports a general need for accounting and budgeting.

Quality of information assets—Simple and complex data formats

Most of the models can handle simple data formats (see Table 21). The LIFE3 model covers five default file types, CMDP provides a list of formats to choose from and PP-CMDS addresses the storage of audio-visual assets. In theory KRDS can be applied to any data types and the DP4Lib model allows cost related to any type of asset to be mapped to sub-services.

	Model									
Types of information assets	T-CMDP	NASA-CET	LIFE3	KRDS	CMDA	CMDP	DP4lib	PP-CMDS	CDL-TPC	EMLTS
Simple data formats	✓	✓	✓	✓	✓	✓	✓	✓	✓	✘
Various simple formats	✓	✓	✓	✓	✓	✓	✓	✘	✘	✘
Complex data formats	✘	✓	✓	✓	✓	✘	✓	✘	✓	✘
Various complex formats	✘	✓	✓	✓	✓	✘	✓	✘	✘	✘

Table 21—Results of the model evaluation for the characteristic Cost variables – information asset types

Most of the models can also handle different types of simple formats. CMDP covers text, email, spreadsheets and databases and the LIFE3 model covers web sites, e-journals, research outputs (theses), sound recordings, and 'other'. CMDA leaves it to the user to determine what complexity means in their context while DP4Lib allows for the mapping of costs related to any type of asset. File types cannot be specified in PP-DMDS and CDL-TCP.

Six of the models can cope with complex data formats. The LIFE3 model can capture quite detailed information for sound recordings and the DP4Lib model can handle any type of asset. PP-CMDS deals with audio and video assets. CDL-TCP does not allow the user to specify data formats and complex formats are not currently present in the pre-defined menus of CMDP but could be built in. All those models which could handle complex data formats could cope with a variety of those. The LIFE3 model covers sound recordings in detail but only allows small databases (up to 10MB) for research data, whilst CMDP covers databases, images and audio formats.

Although complex formats are generally supported it was stated in the stakeholder consultation that one cost model failed to account adequately for the complexity of curated objects. However, this was the only comment stating this and it was addressed to a specific model. Difficulties expressed from stakeholders could indicate problems with the management of complex objects. The requirement of adaptable design of cost models is evaluated in detail in the requirement “Adaptable/modular design”, under the characteristic “Functionality and usability”.

Driver 8:

The ability of cost models to estimate how costs scale with the quality of information assets supports a general need for accounting and budgeting.

Quality of activities

Cost models support different ways to specify the quality of activities. Six models support specification of the quality of activities (see Table 19 above). Four of the models allowed for the unstructured specification of the quality of activities. NASA-CET assumes that principal investigators will assess their confidence in the information provided back by the tool, and KRDS leaves it to the user to determine how to relate costs and quality. LIFE3 covers quality levels associated with digitisation procedures and volume as well as the QA of metadata and policies.

Six of the models allowed for the structured specification of activities. CMDP covers the quality of record repairs as well as comparing the costs of different levels of archive storage. LIFE3 allows the user to indicate the quality of digitisation. CMDA leaves it to the user to determine how to structure quality.

Upload/download capacity of repository system

The NASA-CET model allows defining of activity sets and includes information about "expected number of users" and "estimated average number of requests per user, per year". The LIFE3 model allows the user to refine access in a separate worksheet of the excel spreadsheet and KRDS includes the access frequency as a cost driver. The PP-CMDS tool allows specifying the number of files which are accessed per month. Optionally an adaptive selection of storage systems responsible for the access can be activated to simulate load balancing in order to increase access rates and to mitigate overloading of system resources.

Due to higher frequencies and increased requirements for infrastructure, hardware and software can have an impact on the costs. Higher access rates could reveal bottlenecks of the system's infrastructure if the access rates exceed the available bandwidth. Cost models which do not cover this requirement have more uncertainties in their calculations. Access frequencies are only considered in four of the models (NASA-CET, LIFE3, KRDS &PP-CMDS). The consideration of access frequencies is an identified gap.

Driver 9:

The ability of cost models to estimate how costs scale with the quality of curation activities supports a general need for accounting and budgeting.

Quality of repository

The quality of a repository's system and processes influences costs. Certifications help to establish comparable procedures and quality measurements. Cost models undertaking such certification initiatives may be a way to enable cost comparison across different repositories and systems.

Almost all the evaluated cost models do not include the methods or recommendations of certifications. The KRDS cost model addresses standardisation issues during the phase of evolving preservation functions and file formats (First Mover Innovation). In this phase organisations may need to develop tools, standards and best practices as first innovator. The pre-existence or development of community standards and best practices has major effects on curation costs. There are no explicit recommendations to use certifications, but customised tools and best practices developed by an organisation can evolve to incorporate community standards and be used by similar organisations.

The CMDA defines as prerequisite for the cost model that an organisation using the CMDA model has the philosophy of a trusted digital repository. It does not include compliance with a specific standard or

certification but it assumes compliance to arbitrary standards. The model further states that all costs are related to the quality of the repository.

Driver 10:

The ability of cost models to relate the costs of curation to the quality of a repository, where quality may for example be expressed in terms of certification—degree of trustworthiness—supports the need to compare costs across repositories.

5.2.1.5 Functionality and usability

This characteristic evaluates the functionality and usability of the cost models and their tools. Cost models can be designed for different groups of users with different views and focuses on the costs of curation (account managers and repository managers being cases in point).

Identified gaps

Lack of support for the user group of account managers/department directors

Lack of cost model tools with a GUI

Documentation facilitates the initial application and increases the usability of the models if it is well written, easy to understand and readily available. The time it takes until a user understands the principles of a model (the learning curve) is also part of the usability aspect as well as the tool support. The tool's adaptability, the ease by which it can be customised for individual settings, is also considered in this section. The characteristic consists of ten requirements (see Table 22).

When compared to stakeholders' needs we identified an important lack in the models ability to be used by general managers who are not specialists in digital curation and a lack of tools with a Graphical User Interface (GUI).

	Model									
Functionality and usability	T-CMDP	NASA-CET	LIFE3	KRDS	CMDA	CMDP	DP4lib	PP-CMDS	CDL-TPC	EMLTS
Users—Repository managers	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Users—Account managers	✘	✘	✘	✓	✓	✘	✘	✘	✓	✘
Documentation	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Learning curve	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Efficiency	✓	✓	✓	✘	✘	✓	✓	✓	✓	✓
Tool support	✓	✓	✓	✓	✘	✓	✓	✓	✓	✓
Graphical User Interface	✘	✓	✓	✘	✘	✘	✘	✓	✘	✓
Modular design	✘	✓	✓	✓	✓	✓	✓	✘	✓	✘
Formulas	✓	✓	✓	✘	✓	✓	✓	✘	✓	✓

Table 22—Results of the model evaluation for the characteristic "Functionality and usability"

Intended users—Repository managers (curation specialists)

The cost models are intended for different groups of users—curation specialists as well as non-specialists. In general all the models seem to require some input from repository managers or curation specialists. T-CMDP assumes the existence of a dedicated Archive and Preservation facility with a skilled team, familiar with digital preservation and with IT and record keeping issues. Whilst NASA-CET is intended for

use by principal investigators, a lot of the required information is very technical and would seem to require expert input. KRDS doesn't specify a specific role, but recommends that "Dedicating a person to be responsible for collecting the cost information will save you effort and deliver results of better quality. The person should be responsible for checking the progress of the survey. Use someone who will be seen as independent and trusted by all staff". CMDA is intended for trusted repository staff, CMDP appears to be for archive IT staff and PP-CMDS is intended for the managers of AV archives. The detail at which it is possible to describe different levels of the preservation within DP4Lib makes it suitable for preservation specialists. The intended users of the CDL-TCP and EMLTS are not clear, but they seem to require specialist knowledge.

About one third of the stakeholders stated in the web consultation that repository managers are responsible for accounting and budgeting digital curation in their organisations.

Intended users—Account managers (non-specialists)

KRDS can be used by any role interested in coordinating the cost exercise and the Balanced Score Card approach of the CMDA could make the model more easily applicable to non-specialist users. It would be difficult for non-specialist users to collect and analyse the technical information needed for CMDP and an inability to enter all technical details would reduce the accuracy of the approximations produced by DP4Lib. The modification of OAIS terminology in CDL-TCP in order to expand applicability and increase understanding suggests an intention to include non-specialists.

The stakeholder consultation showed that, in most organisations, non-specialists such as general financial/account managers, department directors and chief executive officers are responsible for accounting and budgeting. Thus there is a clear gap in the models ability to serve non-specialists.

Driver 11:

The ability of cost models to be used by non-specialists in digital curation supports current accounting and budgeting practices and this ability is dependent on good documentation.

Documentation

All the models are accompanied by some documentation. The form of the documentation varies and includes manuals, guides, fact sheets and research papers. There are papers about T-CMDP and there is guidance built into the spreadsheet. There is a large, 95-page user guide and 28-page technical guide for NASA-CET. There is a lot of guidance provided by the LIFE3 model and supporting tools. KRDS has a 46-page user guide including a brief "How-to guide" and there are also a number of fact sheets and flyers to introduce various elements of the framework. There is a research paper about CMDA but not much else and further documentation would increase its usability. Likewise, CMDP could do with further support and guidance to go with the available papers and the guidance within the tool. Whilst documentation, guides and manuals are available for DP4Lib, most of them are only available in German. An overview of the relevant concepts implemented in The PP-CMDS are described in a project deliverable D2.1.2. A 26-page paper describes the workings of CDL-TCP and is clearly presented and makes its workings visible to the user. A paper and a series of blogposts serve to give an overview of the EMLTS, but the model and detailed user instructions for it are unavailable.

The consultation with stakeholders underlined the importance of good and easily understandable documentation for the models. When asked about the possible improvements for cost models, a better definition of activities, better usability and more guidance for the provided formulas was requested. The evaluation also showed that there is a lack of guidance on selection of a model.

Driver 12:

The availability of clearly presented documentation suitable for getting a quick overview of a model as well as in-depth knowledge of its features and assumptions supports the uptake and usability of cost models.

Learning curve

The time needed to understand a cost model depends a great deal upon the user's familiarity with digital curation activities. It is therefore difficult to generally assess it. Descriptions and evaluations of the learning curve are based on the evaluator's knowledge within the domain.

A day would be sufficient for a user to learn to use any one of most of the models, assuming they have the appropriate knowledge. T-CMDP assumes familiarity with the field of digital preservation and CMDA assumes that users are employees of a trusted digital repository. LIFE3 is fairly straightforward but may need input from a digitisation team for specifics. Similarly, if a user can identify the cost types, elements and activities then the DP4Lib model can be applied quickly. One could make a good start with KRDS in a day, but actual calculations of costs would require developing a tool. It may be that the myriad fields and acronyms in NASA-CET are self-explanatory to space science researchers, but they may not be so to other users.

Efficiency

Cost models require certain levels of detailed information as an input in order to be able to compute their results. The evaluators estimated the time needed to collect the necessary information. The model evaluation analysed if the cost models using little input information could provide initial results within an hour or if the models needed more input information resulting in results within half a day.

All cost models except two (KRDS and CMDA) required little input and could be used within an hour. KRDS requires that the organization develop its own calculations based on the user guide.

In the case of T-CMDP this is based on having previously calculated staff time on specific activities. NASA-CET produces quite a lot of detailed information very quickly but analysis may take longer, though this may become quicker with repeated use. A rough idea of costs can be obtained from the LIFE3 model in a relatively short period of time and then fleshed out with more detail. CMDA is designed to be used within an institutional context so local parameters would need to be determined and getting quick results from CMDP requires an understanding of the operational options within the archive. Similarly, the DP4Lib model depends on the available information about the preservation activities. The EMLTS tool is not available, but is described as making use of simulations of the development of storage costs. These simulations are based on the Monte Carlo method, which repeats random sampling of its input variables in order to calculate the probabilities of numeric results.

Tool support and graphical user interface (GUI)

This section assesses the tool implementations for the cost models and also their availability. Different implementations of cost models exist and they range from spreadsheets to online tools. With the exception of CMDA all cost models have some implementation of a tool. T-CMDP uses a spreadsheet as does the NASA-CET cost model. The NASA-CET tool makes use of Microsoft Excel Visual Basic (VBA) and has a graphical user interface. The NASA-CET spreadsheet can add diagrams to the calculated results. The VBA used in NASA-CET can be confusing, with no clear way to proceed from one stage to the next. The LIFE3 model provides a spreadsheet and a web tool and can produce graphical outputs. KRDS has no tool which can be used out of the box and organisations need to implement their own calculations. The CMDA cost model has currently no implemented tool. The CMDP tool is implemented in a spreadsheet with VBA.

The DP4lib model offers a spreadsheet including some default activities as a template for its calculations. PP-CMDS's tool is available as web application and as a Java SE application for Windows, Mac and Linux. The CDL-TCP cost model has a spreadsheet which can be used. The blogposts about the EMLTS show graphical outputs that can be obtained from the model.

The evaluation showed that there is a lack of models with graphical user interfaces.

Driver 13:

The availability of cost tools with attractive and intuitive graphical user interfaces supports the uptake and usability of cost models.

Modular design

Seven of the models as shown in Table 22 above are modular. NASA-CET allows the user to select which elements to include, LIFE3 allows the user to modify elements within the tool quite easily and DP4Lib allows for elements of cost groups and sub-activities to be extended and customised. CMDP is currently missing three modules and yet produces results. KRDS leaves it to the user to describe what to include or remove when they implement their institutional spreadsheet. CMDA is modular in theory. The CDL-TCP model is to some extent modular as it consists of 11 high-level cost categories, which are defined in separate worksheets of its tool implemented as a spreadsheet, and allows to individually change each of these cost categories. This model is also based on a modified version of the OAIS reference model in order to broaden applicability and facilitate understanding.

Formulas

Most of the models use algebraic formulas. Excel formulas are used in T-CMDP, and the LIFE3 model guidance describes the formulas used. CMDP also applies Excel formulas, but more guidance on how they work would be useful. Algebraic formulas were used in the case studies for the CMDA, but access restrictions mean this cannot be specified for the tool itself. NASA-CET uses regression techniques to develop the coefficients for a set of seven trial relationships of FTE to workload parameter for each of the selected workload parameters. A formula to calculate amortisation rates is available in the documentation for the DP4Lib cost model. CDL-TCP provides formulas for the "Pay-as-you-go" and the "Paid-up" price models where the first one assumes an annual billing cycle in contrast to the latter which is based on a one-time payment. Both formulas are clearly explained in the documentation.

Driver 14:

The inclusion of pre-defined and adjustable formulas, parameters and values to guide users on best practices supports the uptake and usability of cost models.

5.2.2 Use case test

In addition to the evaluation of model capabilities, we wanted to test the models against one practical use case. Most of the current models are aimed at organisations whose core mission is preservation. However, as our stakeholder analysis revealed, an increasing number of organisations whose primary function is *not* preservation also need to plan how to retain digital information for the mid to long-term. We developed a simple use case to assess how useful the current models might be for UK Universities wishing to assess in-project and post-project curation costs associated with new research grant applications. We chose this use case as many UK funders will consider requests for resource to cover

research data management and curation costs during the active phase of the project[19]. However, whilst in-project costs may be requested, out-of-project costs must be borne by the institution. Some funders provide subject specific data centres for longer-term retention—for example, the Economic and Social Research Council (ESRC) funds the UK Data Service—but this is not the case for all Research Councils UK (RCUK) funders. As such, many UK higher education institutions (HEIs) are currently struggling to identify in-project curation related costs that can be factored into new grant proposals whilst also grappling to understand the longer term cost implications associated with providing access to research outputs.

The use case describes an application for funding to the Arts and Humanities Research Council (AHRC) for a small digitisation project where 1000 image files would be produced. In the case of AHRC, selected research data outputs must be retained for a period of ten years beyond the end of the project. For the practical evaluation we employed an activity spreadsheet that was developed for the Piloting the LIFE tool in a UK HEI project[20]. The activity spreadsheet is based around the KRDS model.

5.2.2.1 Use case details

Institution Name: University of Glasgow

Number of research staff: 1.2 (Principal Investigator 0.2, Research Assistant 1.0)

Number of technical support staff: 0.5 (Technical Support)

Number of research support staff: 0.3 (Archives staff)

5.2.2.2 Results of use case test

Many of the models we evaluated do cover pre-ingest activity and provide guidance on using activity checklists (for example KRDS, NASA-CET). These activity checklists proved extremely helpful in identifying possible activities to include in cost assessments but for the most part it is left up to the model user to develop their own activity spreadsheets and to develop a means of calculating the costs. Whilst this approach allows for greater institutional control, it is also difficult and time consuming. Based on initial feedback we've had from memory organisations undertaking costs assessment exercises[21], the level of resource required simply to define these parameters may be more than most Universities are currently prepared to commit. As such, it would be beneficial to develop a range of default cost data sets for the academic sector to draw from in the short term to at least estimate costs while more accurate costing approaches are considered.

Where models do provide tools and, in some cases default data, it was often difficult if not impossible to plug the pre-ingest research data management related activity data to into the existing models to assess longer-term preservation costs. As noted above, in-project costs can be requested in new grant applications. However, the University will also need to have an idea about what resource they need to commit for the minimum retention period—in most cases a minimum of 10 years.

Our practical assessment, based on one simple use case, illustrates that costs and benefits models will need to better reflect a much broader array of business types and processes and that improved join up over the lifecycle will be necessary to enable efficient and accurate budgeting[22].

---

[19] This aligns with Research Councils UK's Common Principles on Data Policy, http://www.rcuk.ac.uk/research/datapolicy/ [Accessed 31 Jan 2014]
[20] Piloting the LIFE costs Tool in UK HEIs, http://www.dcc.ac.uk/projects/life [Accessed 31 Jan 2014]
[21] Based on a conversation with Barbara Sierman, KB-NL and Sabine Schrimpf DNB about their experience in undertaking cost assessment relating to curation and preservation.
[22] The EC-funded TIMBUS project is currently exploring the preservation of business processes and workflows. 4C project partners SBA and DPC are also members of the TIMBUS project and we anticipate that the work of this project will feed into the ESRM and roadmap work of 4C.

6 Discussion and Recommendations

In order to navigate within the complex spectrum of cost and benefit models issues, the identification of gaps in current provision is vital for model users and model developers. This section aims to establish a better understanding of the gaps that exist between current models and the stakeholders' needs for financial information; it includes potential incentives for the uptake of the existing models and for future development of cost and benefit models. It focuses on what can be done to ensure that models be developed in accordance with the users' needs, and concludes with a summary of good practices for model developers.

The stakeholder consultation showed that users' primary requirement is for models that are easy to use, reliable and fit for purpose. Overall, the quality of the models reviewed is very high and there are a number of excellent features in each that help specific user communities get a good grasp on their curation costs. But there is room for improvement to make these models more usable and valuable to a wider range of stakeholders. The most challenging of the gaps identified relate to the lack of intuitive and easy to use tool interfaces and simple user-guides; lack of validation of the models within the target communities, inability to adequately model the required use cases; and a lack of standardised definitions of curation to support comparison between alternate options. A last challenge, which is not directly an impediment for the use of the models, is that there is a significant lack of functionality that cater for the users' requirement that models that support both costs and benefits since most of the models are purely cost focussed.

6.1 Usability

Usability

Stakeholders want models that are easy to use

The stakeholder consultation highlighted that the lack of usability is an important gap and one of the barriers that may be limiting take-up of the models. Many of the models reviewed had very detailed user guidance. Few of the models have quick start guides or graphical user interfaces that facilitate usability. Whilst the information contained in the detailed guidance is, no doubt, invaluable as organisations proceed with an in-depth costing analysis, this sort of guidance doesn't really help potential users to determine if a particular model is right for them without expending quite a lot of effort. In many other cases, the guidance is embedded in the tools themselves so users need to start using them to get an idea of how they work and whether they will be right for their particular needs. However, it should be pointed out that some of the reviewed models have enhanced usability by providing pre-defined values and settings. These default settings may serve to help users find out what to consider when doing an assessment of costs, and even provide some guidance on good practices in certain fields.

Recommendation 1:

Make the selection and use of the models easier by developing concise, high-level overviews using plain language and common descriptive elements detailing what the models and tools can provide, who should use them, and when they should be used.

Recommendation 2:

Make the tools more useful by overlaying them with simple graphical user interfaces.

Recommendation 3:

Make the tools more useful by including pre-defined but adjustable parameters for activities (activity checklists), cost elements and variables as well as pre-defined values.

Most of the evaluated models require some input from specialists in digital curation. However, the stakeholder consultation showed that in most cases those responsible for the accounting and budgeting were account managers and general managers. This has important implications for the design of the models, as it implies a need for models that are easy to understand—not relying on in depth technical knowledge for instance—and well described using non-technical language. The models must facilitate communication between curation specialists and non-specialists and this emphasises the importance of a clear documentation of models.

Recommendation 4:

Support the models with clear non-technical documentation to make them easier to understand and use at the management level.

In order to provide accurate results, most of the evaluated models do however require the input of detailed information, and this tends to make the cost models complex to use. Thus, there is always a trade-off between accuracy and the required level of detail.

Recommendation 5:

Design models in a way that allows for rough results with basic information entered, but which will provide increasingly accurate results as more detailed information is added.

6.2 Reliability

Users are inclined to prefer—to trust—models that are based on standards and validated within the community. The accuracy of models also plays an important role. It is however very difficult to evaluate the existing models' accuracy and precision. This is due to the diversity and complexity of the models and because empirical cost data is scarce.

Reliability

Users want accurate, clearly defined, and validated models

Recommendation 6:

Gather cost data sets to test and refine the accuracy and precision of the models.

Reliability is also related to the models' ability to provide clear and transparent definitions of curation costs. Without essential information about what the costs cover the results of cost models have limited utility. The consultation showed that digital curation activities are often part of other business activities which makes it difficult to extract and analyse the costs. For an accurate accounting and budgeting a clearer overview of all activities concerning digital curation is required.

A well understood cost structure would also facilitate outsourcing some or all activities and the estimation of these expenses. Likewise the need to account for benefits and value of curating digital information assets will foster the willingness to separate and breakdown activities.

There is no consensus on how digital curation activities, cost elements and cost variables are termed and defined, and this is a major obstacle for sharing and understanding cost data and contextual information.

To bridge this gap further provision of clear definitions and common understanding of concepts through clear models and controlled vocabularies will remain critical through the remainder of the 4C project and beyond.

Recommendation 7:

Develop and adopt common terms, definitions, controlled vocabularies and concepts related to cost & benefit modelling. Contribute back into the curation community to support standardisation and interoperability of cost models and cost evaluations.

6.3 Adaptability

Adaptability

Users want models that are customisable

As the stakeholder consultation showed, users need to assess the costs and benefits of curating different types of content, in various amounts, various complexities and with different requirements for access and availability. They need the models to fit their purpose—address the required information assets and the right curation activities and resources, and at the right level of detail.

The current models break down costs in different ways, in different activities and functions, in different cost elements (onetime/running; direct/indirect; capital/labour and so on) and in different accounting periods. This is understandable given that most were devised in order to fulfil a particular need in particular organisations with established accounting procedures. However, this adherence to a specific cost breakdown is a barrier to the adoption of a model by organisations that need to handle other amounts and types of information assets and/or apply different curation services. There is a need for a standard costs breakdown, but there is not yet a consensus on how to do this. The stakeholder consultation identified this as a fundamental pre-requisite for driving comparability of curation costing efforts.

6.4 Standardisation

Standardisation

Users want models that allow for comparing costs

The lack of a consistent way to break down costs is a major barrier for comparing and exchanging financial information. For the purpose of streamlining any one organisation and optimizing activities internally, there may not be a requirement for a common standardised breakdown structure. However, if we want to compare costs across organisations or for different services, to learn from each other's practices and identify the most efficient ways of handling digital curation, we need to define and break down costs in a more transparent and uniform way.

Most models break down activities based on the OAIS functional entities, and this seems to be a sound way of describing the activities. However, the OAIS model is an abstract model that intentionally does not reflect actual implementations and practices. The OAIS functional model cannot be directly used as basis for assessing costs since costs can only be assessed for concrete systems and procedures. Another problem with using the OAIS functional breakdown is that some systems may cover more entities or only parts of entities complicating the use of the OAIS model.

Regarding the breakdown of cost in elements and use of general accounting principles there are no standardised ways of doing this within the digital curation community. The Transparent Approach to

Costing (TRAC)[23], which is applied in Higher Education in Britain, has been suggested as a methodology for recording resource cost data [Beagrie et al., 2008, p.13].

The requirement for a standardised way of recording costs is in tension with the fact that at the same time stakeholders also want the models to offer a high degree of flexibility and adaptability to local configurations. This Gordian knot suggests that there is a need to design a high-level cost and benefit framework that can represent most types of organisations and information assets, possibly along the lines of the OAIS standard as well as accounting and budgeting standards. Ideally, this framework should be adjustable to specific use cases.

Recommendation 8:

Reach consensus within the community on how to breakdown costs and describe a framework that at an abstract level models the cost and possibly also the benefits of digital curation in order to promote comparison of alternative scenarios.

To facilitate comparisons of costs we also need more formalised ways of describing the *quality* of the priced curation activities. One way of doing this could be through the means of audit and certification initiatives.

Recommendation 9:

Enable cost models to account for the quality and trustworthiness of curation activities, for example demonstrated through certification.

6.5 Strategic planning

Strategic planning

Users need to be able to assess the benefits

Comparing the costs and benefits of different scenarios to support decision-making and funding requests is a considerable driver for managers, who form the largest potential user group for cost models. Currently only two of the models, CMDA and KRDS, include means of relating costs and benefits, the former using the Balance Score Card (BSC) methodology and the latter using a checklist of benefits (see description of these models and tools in Section 3.3 above).

Awareness of the benefits of curation is essential for organisations—whether they consume or supply curation services—in order for them to sustain their business cases. Whilst the cost of curation basically depends on the quantity and the required quality of the information assets—which, in principle, can be assessed objectively for a particular scenario—the benefits of the scenario depends on the stakeholder perspective—and as such the identification and assessment of benefits is subjective, and this should be reflected in the way that cost and benefit models are designed. To this end the 4C project has also engaged with stakeholders to elicit their priorities regarding various types of benefits, including among others risk, trustworthiness and sustainability, and these concepts and the results of the engagement are described in a deliverable report [4C, D4.1, 2013].

Thus, there is a specific need for associating costs of digital curation with benefits. This is true for both organisations that are looking for monetary profit and for organisations that are trying to elucidate the more intangible, non-financial benefits of digital curation, such as reputation and transparency. It is not

[23] TRAC, http://www.jcpsg.ac.uk/guidance/ [Accessed 31 Jan 2014]

enough to be able to produce accounts and budgets; users also need to express the relation between the investments in digital curation and potential benefits that can be derived. Thus, cost and benefit models for digital curation are important for decision making on alternative solutions and strategic planning, including risk management. Enabling the comparison of the costs and benefits of alternative scenarios is also crucial if organisations want to become more efficient when dealing with digital curation. One of the more obvious needs derived from the stakeholder consultation is this need for efficiency.

The consultation showed that users are generally more interested in using the models for budgeting than for accounting, and that they are generally more concerned with being able to account for running costs than for investment cost. These findings suggest that many organisations are able to get money for one-time investments, but have more trouble justifying the on-going costs, which also explains why they focus on how to plan for future costs rather than past costs.

Recommendation 10:

Link cost models with existing benefit models, or develop models that integrate the assessment of costs and benefits.

Another obstacle hindering uptake appears to be that current models can't easily feed into each other and don't reflect the fact that different stakeholders within an organisation may wish to use different models to calculate costs at different points in time. For example, if a researcher uses NASA-CET or KRDS to calculate in-project, directly incurred costs for a specific project, it is not easy to feed the tailored source data and/or results into another cost model such as LIFE3 or CMDP at the institution-level to help calculate indirect costs associated with longer-term archiving. Conversely, when determining the indirect costs associated with archiving data, some institutions may wish to be able to separate out these costs in instances where a funder provides a data centre for deposit rather than the data being retained by the institution. Cost models need to reflect organisational workflows rather than assume that all cost information will come through a single point (such as an individual or organisational unit). Without being able to assess costs that reflect different workflows involved in generating and caring for data in the short and longer-terms, it will be very difficult to determine potential areas where cost savings and efficiencies can be made.

In addition to being able to exchange information between various models, it would be beneficial if we could reach a point where the models interacted with other organisational systems (business models, HR systems, grant systems). As the models are currently used as stand-alone tools, users would have to manually update the formulas they have developed any time a variable changes (for example salary scales change, VAT increases, and so on). It would be much better if the models were interoperable with other organisational systems that are likely to be more frequently updated and maintained.

Recommendation 11:

Enable the integration of cost and benefit models with external organisational systems thereby allowing them to be semi-automatically updated across parts of the lifecycle.

6.6 Good Practice proposals for model developers

The evaluation of the cost models has led the team to develop the following preliminary insights into what seems to be good practice for developers of models for the cost of digital curation. As in many other areas, the very best practice for model developers is, in general, to keep it simple.

Proposal 1:

Use a standardised definition of digital curation.

Digital curation is not a fully mature profession, and consequently its activities are not clearly defined or agreed upon. Therefore it is difficult to estimate their cost. The lack of a general understanding of the activities of digital curation makes it especially import to adhere to some form of standard, such as OAIS. The OAIS is only a reference standard and does not encompass all areas of cost of digital curation, but it is so far the most detailed and widely known standard for digital curation.

Note, that even an area as well known as basic financial accounting has for decades spent immense resources in defining international standards for accounting (also known as financial reporting).

Proposal 2:

Limit the purpose of the model and define clearly the expected users of the model and its scaling capacity

It is tempting to create an all-inclusive model, but the purpose of the model should be limited to one purpose, such as short term prediction or estimating historical or present cost. A model that tries to satisfy the needs of very different users by modelling very different scenarios of digital curation becomes very complex. Instead, good practice seems to be to limit the user group and the scenarios. As an example trying to model very small or very large organisations in the same model makes the model very complex. Instead, setting narrow lower and upper limits on the model, such as organisation size, amount of material to curate, and diversity of the material makes the model easier to develop, examine and maintain.

Proposal 3:

Favour simplicity; be explicit about limitations on accuracy.

The huge uncertainty in many areas of the cost of digital curation is the main reason why good practice is to favour a clear simply approach, or where granular detail is critical to be explicit about limited guarantees of accuracy. There is simply no point in creating a very complex and detailed model when many areas in general are so uncertain that the model overall may show a very high level of variation.

Proposal 4:

Start out with something rather simple.

Instead of trying to model all activities for digital curation at the beginning, the preferred practise seems to be to start out with more simple and better known activities such as archival storage. If (or when) the simple activities have been modelled the developer can continue on the more complex activities such as ingest or preservation planning.

Proposal 5:

Limit the time scope.

Uncertainty increases with time, and that is even true with digital curation, being dependent on the overall uncertain development of it. It seems good practice to start out with a limited time period (5, 10, 20 years), even though some information assets are expected to be curated for eternity.[24]

Proposal 6:

Use simple formulae.

The cost of the activities of digital curation should be generalised and simplified to such an extent that rather simple formulae can be used. The formulae do not have to be simple in the way that they are linear or very easy to solve, but the need for data in lookup tables to model specific conditions regarding start, end and transitions should be avoided. Generalisations and approximations such as the use of power series/Taylor series should be used, but limited to a few. The formulae should be described in a formal way and explained in text.

Proposal 7:

Implement the model in a simple and widespread tool.

In order to ease use and review of the model a simple and widespread tool such a spreadsheet should be preferred to implementation in a more advanced model language or just a general programming language. Explicit naming of variables and parameters should be used, accompanied with written explanations of the implementation. Example data should be used in the model which the user then can alter.

[24] Only a limited number of stakeholders have a real long-term preservation need. In many cases long term preservation is driven by legal requirements.

7 Conclusions

In this report we have described the work done to evaluate current cost and benefit models in the field of digital curation against stakeholders needs for financial information to reveal gaps in the capabilities of the models and point to ways in which cost and benefit models may be improved to increase their usability.

We first presented a preliminary definition of terms and concepts related to cost and benefit models and placed them within the economic context of digital curation. The components of models have been described, including cost elements and variables. The present work clearly shows that the lack of a universally accepted terminology and clarification of cost and benefit concepts is an important obstacle for reaching consensus on how to model these. Thus this work may serve as a starting point for creating a common understanding of cost and benefit concepts and terminology used within the digital curation community, possibly as integrated in the Economic Sustainability Reference Model (ESRM).

Next we have identified existing cost and benefit models, altogether 10 models of which only two include benefit modules. The scope of each of the models has been described and the core facts about each model presented in tables to allow for easy comparison of the models. The tables present among other things what information assets the models can handle, which activities they cover, which cost elements and variables they account for, and thus provides a summary of the models from which potential users can get a quick overview of what models they may want to use.

Following this we have analysed stakeholders needs for financial information, though interviews with model developers engaged in 4C, from a survey of the needs for financial information of different types of stakeholders as well as through other stakeholder engagement activities conducted by the 4C project.

Based on the analysis of stakeholders' needs we have created a simple evaluation schema where the stakeholder needs have been turned into requirements for models formulated as yes/no questions. This schema has then been used to evaluate all the models. The results of the evaluation have then been used to identify gaps in the models capabilities to meet the stakeholders' needs and to identify good practices for model developers.

One of the most important gaps we identified is the general lack of usability of the models, a gap primarily linked to the documentation associated with the models. Even though the documentation is often very detailed, in most cases a simple introduction to the scope of the model that would allow the potential user to quickly find out if the model is appropriate is a lacking. For this reason we have recommended that creating high-level overall user guides to the models could bridge this gap. Another factor affecting the usability of models is the absence of simple graphical user-interfaces for the tools. Furthermore, we found from evaluating the models that applying pre-defined formulas, parameter and values, could greatly enhance the ease of use of the models and help provide guidance on best practices. One of the clear complaints from stakeholders about models is that they are very complex to use. However, the required level of detail of the models is directly linked to the accuracy of the models and thus represents a trade-off.

Another prominent gap is the lack of the models capability for expressing the quality of digital curation activities and services. The adoption of audit and certification practices will bridge this gap. In addition to the quality issue there is also the question of a lack of the models ability to assess the benefits that the organisations may incur from their spending, a need that was highlighted in the stakeholder consultation.

Finally, the evaluation found that the fact that the models are not easily adaptable to other organisations and assets than the ones they were created for constitutes a serious gap for the uptake of the models. This in turn implies that the models are difficult to validate. Establishing a common method of describing and breaking down costs could bridge this gap, and therefore we have recommended designing a generic cost and benefit framework, maybe based on the OAIS standard and other relevant standards. This model should ideally be customisable to specific scenarios and use-cases. However, it is still an open question whether this is feasible due to the inherent complexity of such all-inclusive cost models. In all circumstances a clearer definition of terms and concepts of cost and benefits of digital curation will provide for a better understanding of these complex relations and facilitate exchange of knowledge about the cost and benefits of digital curation.

References

4C, D2.1—Baseline Study of Stakeholder & Stakeholder Initiatives, 2013, http://4cproject.eu/community-resources/outputs-and-deliverables/d2-1-baseline-study-of-stakeholder-stakeholder-initiatives [Accessed 31 Jan 2014].

4C, MS9—Draft Economic Sustainability Reference Model, 2013, http://4cproject.eu/community-resources/outputs-and-deliverables/ms9-draft-economic-sustainability-reference-model [Accessed 31 Jan 2014].

4C, D4.1—A prioritised assessment of the indirect economic determinants of digital curation, 2013, http://www.4cproject.eu/community-resources/outputs-and-deliverables/d4-1-a-prioritised-assessment-of-the-indirect-economic-determinants-of-digital-curation [Accessed 31 Jan 2014].

APARSEN, D32.1, Report on cost parameters for digital repositories, 2013, http://www.alliancepermanentaccess.org/index.php/knowledge-base/member-resources/documents-and-downloads/?did=123 [Accessed 31 Jan 2014].

APARSEN, D32.2, Report on testing of cost models and further analysis of cost parameters, 2013 http://www.alliancepermanentaccess.org/index.php/knowledge-base/member-resources/documents-and-downloads/?did=150 [Accessed 31 Jan 2014].

Beagrie, N., Chruszcz, J. and Lavoie, B. Keeping Research Data Safe. A Cost Model and Guidance for UK Universities, Copyright HEFCE 2008, http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf [Accessed 31 Jan 2014].

BRT—Blue Ribbon Task Force on Sustainable Digital Preservation and Access, “Sustainable Economics for a Digital Plant: Ensuring Long-Term Access to Digital Information,” Final report, 2010, http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf [Accessed 31 Jan 2014].

CCSDS (Consultative Committee for Space Data Systems), Reference Model for an Open Archival Information System (OAIS), CCSDS 650.0-M-2, Magenta Book, 2012 (ISO14721:2012), http://public.ccsds.org/publications/archive/650x0m2.pdf [Accessed 31 Jan 2014].

CCSDS (Consultative Committee for Space Data Systems), Audit and Certification of Trustworthy Digital Repositories, CCSDS 652.0-M-1, Magenta Book, 2011, (ISO 16363:2012), http://public.ccsds.org/publications/archive/652x0m1.pdf [Accessed 31 Jan 2014].

CCSDS (Consultative Committee for Space Data Systems), Producer-Archive Interface Methodology Abstract Standard (PAIMAS), CCSDS Magenta Book, 2004, (ISO 20652:2006, reviewed and confirmed in 2009) http://public.ccsds.org/publications/archive/651x0m1.pdf [Accessed 31 Jan 2014].

Garrett, J. (co-chair) and D. Waters (co-chair), Preserving Digital Information: Report of the Task Force on Archiving of Digital Information, The Commission on Preservation and Access and The Research Library Group, 1996, http://www.clir.org/pubs/reports/pub63/reports/pub63watersgarrett.pdf [Accessed 31 Jan 2014].

Lunghi, M, N. Grindley, B. Stoklasová, A. Trehub, and C. Egger, “Economic Alignment,” in Aligning National Approaches to Digital Preservation, ed. N. McGovern (Educopia Institute Publication, 2012), 195-268, http://educopia.org/publications/ANADP [Accessed 31 Jan 2014]

Watson, J., The LIFE project research review – Mapping the landscape, riding a life cycle, 2005, http://discovery.ucl.ac.uk/1856/1/review.pdf [Accessed 31 Jan 2014].

External Links

http://4cproject.eu

http://4cproject.eu/community-resources/outputs-and-deliverables/d2-1-baseline-study-of-stakeholder-stakeholder-initiatives

http://4cproject.eu/community-resources/outputs-and-deliverables/ms9-draft-economic-sustainability-reference-model

http://4cproject.eu/community-resources/outputs-and-deliverables/d2-1-baseline-study-of-stakeholder-stakeholder-initiatives

http://www.4cproject.eu/community-resources/outputs-and-deliverables/d4-1-a-prioritised-assessment-of-the-indirect-economic-determinants-of-digital-curation

http://aparsen.digitalpreservation.eu/pub/Main/CostModels/DP4lib-Cost-By-Service-CostModel.docx

http://beagrie.com/krds-i2s2.php

http://blog.dshr.org

http://blog.dshr.org/2011/09/modeling-economics-of-long-term-storage.html

http://blog.dshr.org/2011/11/progress-on-economic-model-of-storage.html

http://blogs.loc.gov/digitalpreservation/2012/06/a-digital-asset-sustainability-and-preservation-cost-bibliography/

http://blogs.loc.gov/digitalpreservation/2012/06/a-digital-asset-sustainability-and-preservation-cost-bibliography/

http://consult.4cproject.eu/index.php/949288/lang-en

http://digitalbevaring.dk

http://discovery.ucl.ac.uk/1856/1/review.pdf

http://discovery.ucl.ac.uk/1856/1/review.pdf

http://dlmforum.typepad.com/Paper_RemcoVerdegem_and_JS_CostModelfordigitalpreservation.pdf

http://dp4lib.langzeitarchivierung.de/downloads/DP4lib-Kostenmodell_eines_LZA-Dienstes_v1.0.pdf

http://dp4lib.langzeitarchivierung.de/index_downloads.php.de

http://educopia.org/publications/ANADP

http://lifedev.hatii.arts.gla.ac.uk/

http://link.springer.com/article/10.1007%2Fs00799-012-0092-1

http://opensource.gsfc.nasa.gov/projects/CET/index.php

http://opensource.gsfc.nasa.gov/projects/CET/index.php

http://opus.bath.ac.uk/32509/

http://prestoprime.it-innovation.soton.ac.uk/imodel/download/

http://PrestoPRIME.it-innovation.soton.ac.uk/planning-tool/accounts/login?next=/planning-tool/

http://public.ccsds.org/publications/archive/650x0m2.pdf

http://public.ccsds.org/publications/archive/651x0m1.pdf

http://public.ccsds.org/publications/archive/652x0m1.pdf

http://wiki.opf-labs.org/display/CDP/Home

http://www.4cproject.eu/news-and-comment/4c-blog/18-call-for-curation-cost-models-by-ulla-bogvad-kejser

http://www.alliancepermanentaccess.org

http://www.alliancepermanentaccess.org/index.php/knowledge-base/member-resources/documents-and-downloads/?did=123

http://www.alliancepermanentaccess.org/index.php/knowledge-base/member-resources/documents-and-downloads/?did=150

http://www.beagrie.com/KeepingResearchDataSafe_UserGuide_v2.pdf

http://www.beagrie.com/krds.php

http://www.clir.org/pubs/reports/pub63/reports/pub63watersgarrett.pdf

http://www.costmodelfordigitalpreservation.dk

http://www.dans.knaw.nl/en/content/categorieen/projecten/costs-digital-archiving-vol-2

http://www.dcc.ac.uk/projects/life

http://www.escholarship.org/uc/item/23b3225n

http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/hole-64.pdf

http://www.ijdc.net/index.php/ijdc/article/view/177

http://www.imaging.org/IST/store/epub.cfm?abstrid=45307

http://www.jcpsg.ac.uk/guidance/

http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf

http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf

http://www.life.ac.uk/

http://www.lockss.org/locksswp/wp-content/uploads/2012/09/unesco2012.pdf

http://www.pv2007.dlr.de/Papers/Fontaine_CostModelObservations.pdf

http://www.rcuk.ac.uk/research/datapolicy/

http://www.tandfonline.com/doi/abs/10.1080/13614576.2010.526014

http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/mow/VC_Kejser_et_al_27_B_1350.pdf

https://prestoprimews.ina.fr/public/deliverables/PP_WP2_D2.1.2_PreservationModellingTools_R0_v1.00.pdf

https://prestoprimews.ina.fr/public/deliverables/PP_WP6_D6.3.1_FM_calculation_R0_v1.01.pdf

https://unsustainableideas.wordpress.com/2011/10/17/update-state-ref-model/

Appendices

A.1 Terms and definitions

In order to contextualise the results of the analysis and evaluation of the cost models it is essential that there is a common understanding of the terms and concepts employed. The key terms used are outlined below.

Term	Definition
Account	A record of past (ex post) financial transactions.
Accounting model	Set of basic assumptions, concepts, principles and procedures that determine the methods of recognizing, recording, measuring and reporting an entity's financial transactions.
Activity	Measureable amount of work performed by systems and/or people to produce a result
Activity checklist	A checklist of digital curation activities that incur costs. Activities may be ordered in categories and different levels of sub categories.
Amortisation	A mechanisms for distributing capital costs over the estimated useful lifetime of an intangible asset to indicate how much of the asset's value has been used.
Benefit model	A representation that describes the benefits (financial and non-financial) and value of digital curation
Budget	A record of estimated future (ex ante) financial transactions over certain periods of time.
Capital cost	Cost incurred once, by acquisition (building space, equipment, materials) or by investments. Also known as investment cost or one-time cost.
Cost element	The units costs can be broken down into (e.g. one-time or recurring costs, capital and labour costs)
Cost model	A representation that describe how resources, such as labour and capital, required for accomplishing digital curation activities relate to costs. Cost models can further be characterised by their cost structure and how they model cost variables.
Cost parameter	A factor that help in defining the system being costed, including cost elements (e.g. capital/labour cost) and cost variables (for example quantity of assets, salary levels).
Cost structure	Defines the way a model break down costs in elements according to the dimensions activity, resource and time.
Cost tool	Implementation of a cost model in an electronic spreadsheet or costing program
Cost variable	Factors that influence the cost, can be divided in service adjustments (for example quantity and quality of assets, and quality of curation services) and in economic adjustments (such as inflation/deflation, depreciation/amortization, and interest [discount rates]).
Digital curation	Digital curation is a series of repository activities including ingest, data management, (archival) storage, preservation planning, access, common services, repository administration, as well as pre-repository (production and pre-ingest (appraisal, selection, preparation, rights), post-repository, and management activities.
Depreciation	A mechanisms for distributing capital costs over the estimated useful lifetime of a tangible asset to indicate how much of the asset's value has been used.
Direct cost	Costs associated with resources used for performing digital curation activities (for example costs of acquisition of storage media, costs of adding metadata), where the amount of resources spent can be directly measured. Also known as variable costs.
Economic model	A representation that describes how economic processes around digital curation work; including the flow of resources (costs and revenues) within the economic lifecycle of digital information assets, and stakeholders (from the demand, supply and management side) interaction with this lifecycle.

Term	Definition
Financial information	All types of information necessary for financial management (accounting and budgeting). It includes factual data on the costs (for example labour, materials and overhead), additional information describing what is being costed (such as assumptions and specifications), as well as information that relates to the benefits and value that the digital curation activities accrue and how these incentives influence economic behaviour and performance.
Fixed cost	Costs, which do not vary with the amount of production. Often the same as indirect costs.
Full- time equivalent (FTE)	A unit that indicates the workload of a worker by expressing the ratio of the total number of paid hours during a period by the number of working hours in that period. Also known as annual work unit (AWU). Used to make workloads comparable.
Indirect cost	Costs incurred by the usage of shared resources, such as general management and administration or common facilities and systems, where it has not been possible to distribute the cost on specific activities. Also known as residual cost or overhead.
Investment cost	See capital cost.
Labour cost	Cost of wages paid to workers.
One-time cost	See capital cost.
Operating cost	See recurring cost.
Periodic cost	Cost that are repeated and incur at intervals (for example some licenses). Also known as term cost.
Recurring cost	On going cost (such as from consumption of media, energy and labour). Also known as running cost or operating cost.
Resource cost	Cost associated with a particular type of resource, capital or labour.
Running cost	See recurring cost.
Stakeholder	On the one side the roles of managers and administrators of digital repositories and other suppliers of curation services; and on the other side the roles of owners, producers and consumers (beneficiaries) of digital assets that have a demand for these services and a willingness to pay for the value that they represent.
Value	Numerical, Boolean, ordered lists that are assigned to a parameter, or the result of a function.
Variable cost	Costs, which vary directly with the amount of production. Often the same as direct costs.

Table 23—Terms and definitions

A.2 Questionnaire for stakeholders

These are the questions from the stakeholder consultation. The full list of choices available for each question can be found in 4C deliverable D2.1 [4C, D2.1, 2013].

Q1. What is the description that best fits your organisation?

Q2. In which country does your organisation reside?

Q3. What is your organisation's core business activity?

Q4. Who are the users or customers of your organisation?

Q5. What are the main funding sources for your organisation?

Q6. What is the global annual budget for your organisation?

Q8. Would you be willing to share curation cost information under confidential conditions with the 4C project?

Q9. Under what conditions would you be willing to share cost information with a wider community?

Q10. Are you interested in being contacted by the 4C project in the next couple of months for further questions and engagement activities?

Q11. Would you like to be informed about future activities of the 4C project?

Q12. Please leave your name and email for further contact.

Q13. Would you be willing to answer some additional questions about your digital curation activities, accounting and cost modelling? There are up to 33 extra questions and will take you on average 25 minutes.

Q14. What are the main funding sources for your digital curation activities?

Q15. What is the annual budget for your digital curation activities (including operational costs)?

Q16. How important are the digital curation activities when compared with your other business activities?

Q17. Are the digital curation activities performed in-house or outsourced? Please specify the pricing model in the comment.

Q18. What infrastructure does your organisation use for digital curation? Please give details in comment.

Q19. How often are assets accessed by consumers?

Q20. How does your organisation currently breakdown the costs of digital curation activities?

Q21. What types of information assets do you need to curate?

Q22. What is the motivation for keeping these assets?

Q23. Who are the producers of the assets for curation?

Q24. Who are the consumers of the assets curated?

Q25. What benefits do the assets represent to your organisation?

Q26. Over what timescales does your organisation need to maintain access to the assets?

Q27. What is the current volume and the projected yearly increase for the next 5 years of the assets kept by your organisation?

Q28. For what purposes does your organisation need financial information related to digital curation?

Q29. Who is responsible for accounting and budgeting for digital curation in your organisation?

Q30. How do you determine the costs of curation in your organisation?

Q31. How often does your organisation need to prepare accounts and budgets for digital curation?

Q32. What type of costs does your organisation need to account for?

Q33. How do you think your organization is likely to benefit from digital curation cost modelling?

Q34. Select the 3 main reasons for your organisation to use a cost model.

Q35. On what basis would you select a cost model?

Q36. Have you ever tried a cost model for digital curation?

Q37. Was the cost model effective?

Q38. How could the cost model be improved?

Q39. To your knowledge, is this model used by other organisations?

Q40. What is the origin of the cost model?

Q41. Does the model cover the activities required by your organisation in the right grouping and the right level?

Q42. Is this digital curation cost model integrated with other cost models used by your organisation?

Q43. What features does the model include?

Q44. Do you have any request for additional features to the cost model?

A.3 List of Stakeholders' Needs

Needs identified from the web consultation, grouped by sections of the consultation.

Activities

- Need for accuracy in budgeting
- Need to understand the significance of the amount and frequency of uploads (ingest) and downloads (access) on the costs
- Knowledge of the costs of curating digital assets
- Knowledge if the models' breakdown structure is based on a standard
- Knowledge if the models' structure is locally defined
- Overview of functional entities within digital curation that incur costs

Content

- Need to know if the model can handle the required type of assets
- Need to account for benefits and value of curating digital information assets
- Need to ensure sustainability of digital assets
- Need to know the relation between the quantity of assets and the costs

Accounting & budgeting

- Need to ensure financing
- Need to ensure annual funding
- Need to ensure that indirect costs are accounted for
- Need for hiring the right types of staff
- Need to make financial adjustments
- Need to have an overview of activities in digital curation that incur costs (Activity checklist)
- Need to prepare accounts for legal or business purposes
- Need to prepare budgets for legal or business purposes
- Need to prepare budgets over a certain time scale
- Need for the model to be usable for the right staff (Intended users)
- Need to trust the model and know the principles it builds on (documentation)
- Need for adaptable (extensible) tool
- Need for clear definitions
- Need for maintainable tool
- Need for guidance / best practice
- Need to obtain the results of a cost assessment within reasonable time

Cost modelling

- Need to know the cost of curating digital assets (Cost of curating digital assets)
- Need to know the relation between the costs and the quality of the activities; to compare alternative solutions and select the most efficient ones (Quality of activities)
- Need to ensure sustainability of digital assets (economic lifecycle of digital information assets)
- Need for enhancing the efficiency and ease of use of the model (GUI)
- Need to obtain cost figures in an automatic and consistent way
- Need to be able to apply the model within reasonable time (learning curve)
- Level of details (OAIS structure)
- Need to accommodate the model to specific scenarios (use cases)

A.4 Condensed Model Evaluation Schema

		Model										
ID	Requirement	T-CMDP	NASA-CET	LIFE3	KRDS	CMDA	CMDP	DP4lib	PP-CMDS	CDL-TPC	EMLTS	ESRM
Model type												
1	Economic model?	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✓
2	Benefit model?	✘	✘	✘	✓	✓	✘	✘	✘	✘	✘	n.a.
3	Benefits-unstructured?	✘	✘	✘	✓	✓	✘	✘	✘	✘	✘	n.a.
4	Benefits-structured?	✘	✘	✘	✓	✓	✘	✘	✘	✘	✘	n.a.
5	Cost model?	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	n.a.
6	Activity checklist?	✓	✓	✓	✓	✓	✓	✓	✘	✘	✘	n.a.
7	Past/current costs?	✓	✓	✓	✓	✓	✓	✓	✘	✘	✘	n.a.
8	Future costs?	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	n.a.
9	Future costs - short term	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	n.a.
10	Future costs-medium term	✓	✓	✓	✓	✘	✓	✘	✓	✓	✓	n.a.
11	Future costs-long term	✓	✓	✓	✓	✘	✘	✘	✓	✓	✓	n.a.
Ressource breakdown												
12	Direct capital costs?	✓	✓	✓	✓	✓	✓	✓	✘	✓	✓	n.a.
13	Direct maintenance/operation costs?	✓	✓	✓	✓	✓	✓	✓	✘	✓	✓	n.a.
14	Indirect costs?	✓	✓	✓	✓	✓	✘	✓	✘	✓	✘	n.a.
15	Labour costs?	✓	✓	✓	✓	✓	✓	✓	✘	✓	✘	n.a.
16	Differentiate labour costs?	✓	✓	✓	✓	✓	✓	✓	✘	✓	✘	n.a.
17	Depreciation/amortization?	✓	✓	✓	✓	✘	✓	✓	✓	✓	✓	n.a.
18	Inflation/deflation (discounting)?	✘	✓	✓	✓	✘	✘	✘	✘	✓	✓	n.a.
Activity breakdown												
19	Standardized breakdown?	✓	✓	✓	✓	✓	✓	✓	✘	✓	✘	n.a.
20	Custom breakdown?	✓	✓	✓	✓	✓	✓	✘	✓	✓	✓	n.a.
21	Breakdown by amount?	✓	✓	✓	✓	✓	✓	✓	✓	✓	✘	n.a.
22	Breakdown by department?	✓	✓	✓	✓	✓	✘	✘	✘	✘	✘	n.a.
23	Breakdown by activity?	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	n.a.
24	Breakdown by other criteria?	✓	✓	✓	✓	✓	✓	✓	✘	✘	✘	n.a.
25	Breakdown by OAIS entities?	✓	✓	✓	✓	✓	✓	✓	✘	✓	✘	n.a.
26	Pre-ingest?	✓	✓	✓	✓	✓	✓	✘	✘	✘	✘	n.a.
27	Pre-Ingest, further details?	✘	✓	✓	✓	✓	✘	✘	✘	✘	✘	n.a.
28	Ingest?	✓	✓	✓	✓	✓	✓	✓	✘	✘	✘	n.a.

		Model										
ID	Requirement	T-CMDP	NASA-CET	LIFE3	KRDS	CMDA	CMDP	DP4lib	PP-CMDS	CDL-TPC	EMLTS	ESRM
29	Ingest, further details?	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	n.a.
30	Data Management?	✓	✓	✓	✓	✓	✗	✓	✗	✗	✗	n.a.
31	Data Management, further details?	✗	✓	✓	✗	✓	✗	✓	✗	✗	✗	n.a.
32	Archival Storage?	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	n.a.
33	Archival Storage, further details?	✗	✓	✓	✓	✓	✓	✓	✗	✗	✗	n.a.
34	Preservation Planning?	✓	✗	✓	✓	✓	✓	✓	✗	✗	✗	n.a.
35	Preservation Planning, further details?	✗	✗	✗	✓	✓	✓	✓	✗	✗	✗	n.a.
36	Access?	✗	✓	✓	✓	✓	✗	✓	✗	✗	✗	n.a.
37	Access, further details?	✗	✓	✓	✓	✓	✗	✓	✗	✗	✗	n.a.
38	Administration?	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	n.a.
39	Administration, further details?	✗	✓	✓	✗	✗	✓	✓	✗	✗	✗	n.a.
40	Common Services?	✗	✓	✓	✓	✓	✗	✓	✗	✗	✗	n.a.
41	Production?	✗	✓	✓	✓	✓	✗	✗	✗	✗	✗	n.a.
42	Management?	✓	✓	✓	✓	✓	✗	✓	✗	✓	✗	n.a.
43	Other activities?	✗	✓	✓	✗	✓	✗	✗	✗	✓	✗	n.a.
Cost variables												
44	Number of assets?	✓	✓	✓	✓	✓	✓	✗	✓	✗	✗	n.a.
45	Volume of assets?	✓	✓	✓	✓	✗	✓	✓	✓	✓	✗	n.a.
46	Yearly increase (number)?	✗	✓	✓	✓	✓	✓	✗	✓	✗	✗	n.a.
47	Yearly increase (volume)?	✗	✓	✓	✓	✗	✓	✗	✓	✗	✗	n.a.
48	Upload/download capacity?	✗	✓	✓	✓	✗	✗	✗	✓	✗	✗	n.a.
49	Simple data formats?	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	n.a.
50	Different simple formats?	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	n.a.
51	Complex data formats?	✗	✓	✓	✓	✓	✗	✓	✗	✓	✗	n.a.
52	Different complex formats?	✗	✓	✓	✓	✓	✗	✓	✗	✗	✗	n.a.
53	Minimum amount of assets?	✓	✗	✗	✓	✓	✗	✓	✓	✗	✓	n.a.
54	Maximum amount of assets?	✓	✓	✗	✓	✓	✗	✓	✓	✗	✓	n.a.
55	Migration strategy?	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	n.a.
56	Migration on demand strategy?	✗	✗	✗	✓	✓	✓	✓	✓	✗	✗	n.a.
57	Migration with normalisation strategy?	✓	✗	✓	✓	✓	✓	✓	✗	✗	✗	n.a.
58	Emulation strategy?	✓	✗	✓	✓	✓	✗	✓	✗	✗	✗	n.a.
59	Quality of activities, unstructured?	✗	✓	✓	✓	✓	✗	✗	✗	✗	✗	n.a.

		Model										
ID	Requirement	T-CMDP	NASA-CET	LIFE3	KRDS	CMDA	CMDP	DP4lib	PP-CMDS	CDL-TPC	EMLTS	ESRM
60	Quality of activities, structured?	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	n.a.
61	Quality of repository, structured?	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	n.a.
Usability of tool												
62	Intended users, repository managers?	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	n.a.
63	Intended users, account managers?	✗	✗	✗	✓	✓	✗	✗	✗	✓	✗	n.a.
64	Well documented?	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	n.a.
65	Learning curve, <1day?	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	n.a.
66	Learning curve, <1week?	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	n.a.
67	Learning curve, <2week?	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	n.a.
68	Get results, 1 h?	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓	n.a.
69	Get results, 1/2 d?	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓	n.a.
70	Implemented in spreadsheet?	✓	✓	✓	✓	✗	✓	✓	✗	✓	✗	n.a.
71	Implemented in web app?	✗	✗	✓	✗	✗	✗	✗	✓	✗	✗	n.a.
72	Implemented in other app?	✗	✗	✓	✗	✗	✗	✗	✓	✗	✓	n.a.
73	GUI?	✗	✓	✓	✗	✗	✗	✗	✓	✗	✓	n.a.
74	Modular?	✗	✓	✓	✓	✓	✓	✓	✗	✓	✗	n.a.
75	Algebra formulas?	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	n.a.
76	Design well documented?	✓	✓	✓	✗	✗	✓	✓	✓	✓	✗	n.a.
77	Pre-set values?	✓	✓	✓	✗	✗	✓	✗	✓	✓	✓	n.a.
78	Parameters/values changable?	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	n.a.
79	Pre-set values for more use cases?	✗	✗	✓	✗	✗	✗	✗	✓	✓	✗	n.a.

n.a. = not applicable

A.5 Inventory of Digital Preservation Solutions

There are many systems available that could be described as occupying the digital preservation solution niche[25]. This appendix is by no means exhaustive, but it gives a flavour of the systems—both commercial, open source and hybrid—that are currently available. Additional and/or updated information may be published when an on-line version of this deliverable is made available on the 4C website.

Company/ Organisation	Arkivum
Product	A-Stor
Description	• Managed digital archiving service • Three copy strategy (two in geographically separated data centres and one offline) • Territory specific data centres (safe harbour data protection) • Local gateway appliances • Encrypted • API • Based on open standards • Cloud based (public or private—OSCAR (On-site Cloud ARchive)) • Offers "100% certain data safety" • ISO27001 • Commercial
Further info	arkivum.com/

Company/ Organisation	Artefactual
Product	Archivematica.
Description	• Standards compliant (OAIS and others) • Micro-service approach • Web dashboard • API • Server (as opposed to cloud) based • Open source
Further info	http://www.archivematica.org

[25] Or at least providing much of the functionality ascribed to digital preservation systems.. In many cases it can be difficult to decide where, for instance, a content management system or a literature database or a repository ends and a preservation system begins.

Company/ Organisation	Atos Origin
Product	SPAR (Système de Préservation et d'Archivage Réparti – Distributed Preservation and Archiving System)
Description	• Independent module system based upon the OAIS model • Created with open source tools for the French National Library • Standards based (Dublin core) • Distributed system • Multiple copies and continuous monitoring • Built in format migration as required • Commercial
Further info	http://www.bnf.fr/en/preservation_spar_old/s.preservation_SPAR_presentation_old.html?first_Art=oui

Company/ Organisation	bepress
Product	Digital Commons
Description	• Hosted institutional repository software • Dedicated discovery network (Digital Commons Network) • "SelectedWorks" individual scholar pages • Web based (customisable front end) • Multi-format • Supports LOCKSS • Standards compliant • Built-in peer review tools • On the fly format conversion on ingest • Commercial
Further info	http://digitalcommons.bepress.com/

Company/ Organisation	BioMed Central
Product	Open Repository
Description	• Hosted solution • Customised DSpace repositories • Standards based • Built on Open source software (DSpace) • (see entry for DSpace for further information) • Commercial
Further info	http://www.openrepository.com/

Company/ Organisation	Duraspace
Product	DSpace
Description	• Standards compliant—including Dublin core metadata schema and Open Archives Imitative Protocol (OAI-PMH) • Supports multiple digital formats • Designed for multi-disciplinary use • Requires persistent identifiers • Tomcat Server (as opposed to cloud) based (although virtualised cloud hosting is possible) • Multiple associated import/export tools • Active user community • Java APIs • Open source
Further info	http://www.dspace.org/

Company/ Organisation	DuraSpace
Product	DuraCloud
Description	• Managed multiple copy / multi-cloud based preservation • Web dashboard • Continuous content integrity monitoring • Built in end user tools including media streaming • Designed to be flexible/scalable on demand • Hosted (cloud based) • Commercial
Further info	http://www.duracloud.org/

Company/ Organisation	Duraspace
Product	Fedora.
Description	• Multi format—content and metadata • Scalable • Web APIs (REST/SOAP) • RDF search (SPARQL) • Content Model Architecture (define "types" of objects by their content) • Multiple storage options • JMS messaging • Web GUI • Open source
Further info	http://www.fedora-commons.org/

Company/ Organisation	EPrints
Product	EPrints
Description	• OAI compliant open repository software • Optimised for Google scholar • Forms the backbone of a number of commercial/hosted repository services • Conform with research funder open access mandates • Standards compliant • RSS feeds • Multi-format deposits • Open source • Commercial managed service available
Further info	http://www.eprints.org/

Company/ Organisation	Ex-Libris
Product	Ex Libris Rosetta.
Description	• Distributed architecture (scalable) • Separate working repository and permanent repository • Standards based (OAIS) • Trusted digital Repository (TDR) Compliant • Web interface • Application programming interface (API) and Software development kit (SDK) available • Commercial
Further info	http://www.exlibrisgroup.com/category/DigitalHeritage

Company/ Organisation	Intrallect
Product	intraLibrary
Description	• Multi format file repository • Part of an Integrated Learning Environment • Standards compliant • User personalisation • Accessed via Web interface • All objects directly addressable (via URLs) • Commercial
Further info	http://www.intrallect.com/solutions/managing_content/

Company/ Organisation	Invenio (CERN)
Product	Invenio.
Description	• Web based document repository • Navigable collection tree • Built in search engine based around specially designed indexes • Flexible metadata • User personalisation and collaboration features • Multiple output formats • Standards compliant • Scalable—split web and database servers • Server based (as opposed to cloud based) • Open source
Further info	http://invenio-software.org/

Company/ Organisation	LOCKSS
Product	LOCKSS.(Lots of Copies Keep Stuff Safe)
Description	• Provides access to purchased digital content whenever the source site is unavailable • Distributed local access through digital preservation appliances (LOCKSS Box) • Peer to Peer monitoring and content repair throughout the LOCKSS network • Standards compliant • Web admin interface • Built in audit and verification tools • Content preserved in original format and migrated on access if necessary • Server based (as opposed to cloud based) • Open source
Further info	http://www.lockss.org/

Company/ Organisation	Microsoft Research
Product	Research-Output Repository Platform - Zenity
Description	• Discontinued April 2013
Further info	http://research.microsoft.com/en-us/projects/zentity/

Company/ Organisation	OCLC
Product	CONTENTdm
Description	• Hosted service with self branding for website, and/or • Local service on Linux or Windows platforms • WorldCat meta data exposure built in • Multi format data storage with customisable metadata • Optimised for customisation via an API • Scalable • Standards based (including Qualified Dublin Core) • Commercial
Further info	http://www.oclc.org/contentdm

Company/ Organisation	Pearson
Product	Equella
Description	• Hybrid/combined leaching and learning, research, media and library content repository • API toolkit to enable Learning Management System (LMS) integration • Hosted solutions available • Commercial
Further info	http://www.equella.com/

Company/ Organisation	**Portico**
Product	Portico Digital Preservation Service
Description	• Dark archive • Specialist preservation of digital publications • Uses format-based migration strategy • Provides perpetual access mechanism—access to content after it is no longer available from other sources • Not-for-profit
Further info	http://www.portico.org/

Company/ Organisation	Tessella
Product	Preservica.
Description	• Cloud and on premise versions available • On site version—single or multi server • Scalable across multiple servers • Cloud version—Amazon S3 and/or Glacier Cloud storage • OAIS compliant workflows • Public access/discovery • CALM catalogue synchronization • Linked Data Registries • Commercial
Further info	http://preservica.com/

Company/ Organisation	Tessella
Product	Safety Deposit Box (SDB)
Description	• Digital Archive platform that provides “Active Preservation” Capabilities • Based on OAIS reference model • Ingest toolkit • Uses risk based approach to preservation planning • Offered in 3 primary configurations: Standalone; Black-box Archive, Active preservation plugin • Commercial
Further info	http://tessella.com/products/tessella-sdb/

Company/ Organisation	VTLS
Product	Vital
Description	• Digital asset management solution • Configurable web interface • Standards compliant (Dublin Core, etc) • Modifiable metadata format • Automatic metadata capture • Server (as opposed to cloud) based • Also available as Software as a Service (SaaS) • Open source core • Commercial
Further info	http://www.vtls.com/products/vital

Digital Preservation Initiatives

As with Digital Preservation Solutions there are many Digital Preservation Initiatives and initiatives with indirect digital preservation connections. Hence this is not an exhaustive list. Some initiatives are also solution providers and have been presented in the tables above.

Company/ Organisation	Description	Further Info
APARSEN	"…to bring coherence, cohesion and continuity to research into barriers to the long-term accessibility and usability of digital information and data… ..building a long-lived Virtual Centre of Digital Preservation Excellence."	http://www.alliancepermanentaccess.org/
ARKive	"…an awe-inspiring record of life on Earth" focusing in particular on imagery of endangered species	http://www.arkive.org/
Blue Ribbon Task Force on Sustainable Digital Preservation and Access	"…principles and actions to support long-term economic sustainability [of digital information]"	http://brtf.sdsc.edu/index.html
Cost Forecasting Model For New Digitization Projects	Specialist modelling of the costs of digitising book collections	http://www.cni.org/topics/digital-libraries/cost-forecasting-model/
Cost Model for Digital Preservation (CMDP)	"…to develop a tool that calculates present and future costs of cultural heritage institutions' digital collections based on various user inputs"	http://www.costmodelfordigitalpreservation.dk/
Costs of Digital Archiving vol. 2	"This project aims at generating a cost model for archiving and disseminating digital scholarly datasets"	http://www.dans.knaw.nl/en/content/categorieen/projecten/costs-digital-archiving-vol-2
Digital Curation Centre (DCC)	"…a world-leading centre of expertise in digital information curation with a focus on building capacity, capability and skills for research data management"	http://www.dcc.ac.uk/
Digital Preservation Coalition (DPC)	"...an advocate and catalyst for digital preservation, enabling our members to deliver resilient long-term access to content and services, and helping them derive enduring value from digital collections."	http://dpconline.org/
Digital Record Object Identifier (DROID)	File identification tool using digital "fingerprint" to identify file formats	http://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/
DP4lib	"…a long-term preservation infrastructure which offers maximum usability and flexibility."	http://dp4lib.langzeitarchivierung.de/ (German only)
DPN (Digital Preservation Network)	"…formed to ensure that the complete scholarly record is preserved for future generations."	http://www.dpn.org/
Elsevier Science digital archive	"…all ScienceDirect Elsevier journals and book series content [in a] digital dark archive "e-Depot" managed by the Koninklijke Bibliotheek (National Library of the Netherlands)	http://http//www.kb.nl/

Company/ Organisation	Description	Further Info
ENSURE (Enabling kNowledge Sustainability and Recovery for Economic Value)	"…[addressing the problem of] spiraling amounts of data produced or controlled by organizations with commercial interests…"	http://ensure-fp7-plone.fe.up.pt/site/
FDsys (Federal Digital System)	"…free online access to official publications from all three branches of the [US] Federal Government."	http://www.gpo.gov/fdsys/search/home.action
Hewlett Packard	"HP Labs is working to create affordable ways to detect and repair damage to data, concentrating on developing high-level software processes"	http://www.hpl.hp.com/research/about/preservation.html
International Dunhuang Project	"…international collaboration to make information and images of all manuscripts, paintings, textiles and artefacts from Dunhuang and archaeological sites of the Eastern Silk Road freely available on the Internet"	http://idp.bl.uk/
Internet Archive	"…a digital library of Internet sites and other cultural artifacts in digital form."	https://archive.org/
interPARES (International Research on Permanent Authentic Records in Electronic Systems)	"…developing the knowledge essential to the long-term preservation of authentic records created and/or maintained in digital form and providing the basis for standards, policies, strategies and plans of action…"	http://www.interpares.org/
Keeping Emulation Environments Portable	"The overall aim of the project is to facilitate universal access to our cultural heritage by developing flexible tools for accessing, manipulating and storing a wide range of digital objects using emulation tools either to reproduce the original environment in which they were created or to enable those objects to be migrated accurately to another environment."	http://www.keep-project.eu/ezpub2/index.php
Keeping Research Data Safe	"…cost/benefit studies, tools and methodologies that focus on the challenges of assessing costs and benefits of curation and preservation of research data."	http://www.beagrie.com/krds.php
LIFE (Life Cycle Information for E-Literature)	"…a methodology to model the digital lifecycle and calculate the costs of preserving digital information for the next 5, 10 or 20 years."	http://www.life.ac.uk/
MetaArchive Cooperative	"…a community-owned, community-led initiative comprised of libraries, archives, and other digital memory organizations…. … a secure and cost-effective repository that provides for the long-term care of digital materials"	http://www.metaarchive.org/
Microsoft Research	"The eHeritage projects preserve cultural heritage through the application of advanced computing technologies"	http://research.microsoft.com/en-us/collaboration/global/asia-pacific/programs/eheritage-projects.aspx
National Digital Heritage Archive	"…a partnership between the National Library, Ex Libris and Sun Microsystems"	http://natlib.govt.nz/about-us
National Digital Information Infrastructure and Preservation Program	"…a national strategy to collect, preserve and make available significant digital content, especially information that is created in digital form only, for current and future generations."	http://www.digitalpreservation.gov/

Company/ Organisation	Description	Further Info
PADI (Preserving Access to Digital Information)	"…aims to provide mechanisms that will help to ensure that information in digital form is managed with appropriate consideration for preservation and future access." (now archived)	http:// www.nla.gov.au/padi/ (http://pandora.nla.gov.au/pan/10691/20110824-1153/www.nla.gov.au/padi/index.html)
Preservation and Long-term Access through Networked Services (PLANETS)	"The primary goal for Planets is to build practical services and tools to help ensure long-term access to our digital cultural and scientific assets."	http://www.planets-project.eu/
PrestoPRIME	"…will research and develop practical solutions for the long-term preservation of digital media objects, programmes and collections, and find ways to increase access by integrating the media archives with European on-line digital libraries in a digital preservation framework."	http://www.prestoprime.org/
PRONOM	Online file format registry	http://apps.nationalarchives.gov.uk/PRONOM/Default.aspx
Shapell Manuscript Foundation	"…dedicated to the collection and research of original manuscripts and historical documents. The Foundation's focus is on the histories of the United States and the Holy Land, with emphasis on the 19th and 20th centuries"	http://www.shapell.org/
TCP (Total cost of Preservation)	"…an analytical framework for modeling the full economic costs of preservation"	https://wiki.ucop.edu/display/Curation/Cost+Modeling
WikiTeam	"…a group dedicated to preserving digital history … … Its primary focus is the copying and preservation of content housed by at-risk services."	http://www.archiveteam.org/index.php?title=Main_Page
IBM	Long Term Digital Preservation (LTDP) projects designed to address the problem of keeping digital information so that the same information can be used at some point in the future in spite of obsolescence of everything involved	https://www.research.ibm.com/haifa/projects/storage/ltdp/index.shtml
Canadian Government Information Private LOCKSS Network (CGI PLN)	"The mission of the CGI PLN is to preserve digital collections of government information."	http://plnwiki.lockss.org/wiki/index.php/CGI_network