Collaboration to Clarify the Cost of Curation

D3.3—Curation Costs Exchange Framework

Deliverable Lead:	The Danish National Archives
Related Work package:	WP3
Author(s):	Alex Thirifays [Danish National Archives], Diana Sisu [University of Edinburgh], Joy Davidson [University of Glasgow], Katarina Haage [Deutsche Nationalbibliothek], Luis Faria [KEEPS], Marjan Grootveld (DANS), Paul Stokes [Jisc], Sarah Middleton ([Digital Preservation Coalition]
Dissemination level:	Public
Submission date:	31 October 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	30 Sep 2014	First draft	Alex Thirifays
0.02	9 Oct 2014	Content review 1 and provision of section 5 and 6	Sarah Middleton
0.03	15 Oct 2014	Content review 2	Ulla Bøgvad Kejser
0.04	15 Oct 2014	Content review 3	Diana Sisu
0.05	17 Oct 2014	Content review 4	Hervé L'Hours
0.06	17 Oct 2014	Provision of section 1.5	Joy Davidson
0.07	18 Oct 2014	Provision of section 3.3	Luis Faria
0.08	20 Oct 2014	Content review 5	Hervé L'Hours
0.09	21 Oct 2014	Content review 6	Diana Sisu
0.10	22 Oct 2014	Provision of section 4	Paul Stokes
0.11	22 Oct 2014	Consolidation	Alex Thirifays
0.12	25 Oct 2014	Final review	Paul Stokes
1.00	31 Oct 2014	Released version	Paul Stokes

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, Nationalbibliotek Og Kobenhavns Universitetsbibliotek	KBDK	DK
Instituto de Engenharia de Sistemas e Computadores, Investigacao e Desenvolvimento em Lisboa	INESC-ID	PT
Danish National Archives	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

Author(s):	Alex Thirifays [DNA], Diana Sisu [UEDIN-DCC], Hervé L'Hours (UESSEX), Joy Davidson [HATII-DCC], Katarina Haage [DNB], Luis Faria [KEEPS], Marjan Grootveld (DANS), Paul Stokes [JISC], Sarah Middleton ([DPC], Ulla Bøgvad Kejser [KBDK]

Table of Contents

Figures

Tables

Executive Summary

To date there have been very few opportunities (and—due to trust issues—little willingness) for digital curation practitioners to exchange information relating to the cost of curation. The Cost Comparison Tool (CCT), the mainstay of the Curation Costs Exchange (CCEx) platform (http://www.curationexchange.org), allows users to upload their curation costs and compare them with those of others in a secure, controlled fashion. Transparency of digital curation costs will help organisations identify greater efficiencies and pinpoint potential optimisations. Insight into how and why peers target their investments can lead to better use of resources, help identify weaknesses and drivers in current practices, and inspire innovations. Last but not least, a community such as that facilitated by the CCEx enables members to collaborate more, communicate more, exchange information and experiences and start addressing the taboo subject of sharing financial information.

This report along with the CCEx, the CCT (http://www.curationexchange.org/compare-costs) together make up the deliverable (D3.3—Curation Costs Exchange Framework) and are intended to address the problems outlined above.

Section 1—Introduction clarifies the purpose, the scope, the terminology, the prerequisites, the audience and the benefits of the Cost Comparison Tool.

Section 2—Description of the Cost Comparison Tool describes the step-by-step process of submitting cost data to the tool and indicates which results are displayed.

Section 3—Rationale and mechanisms outlines the approach selected and elaborates on the reasoning behind the most important principles of the CCT. The section also explains how the calculations are made.

Persuading users to use the tool and submit cost data has proved to be a difficult task. We have invested significant resource into addressing security, anonymity and confidentiality issues in order to reassure stakeholders who are considering sharing sensitive data. This topic is addressed in **section 4—Confidentiality and anonymity concerning the submitted data.**

We have also invested effort into communicating the benefits of sharing cost data and of using the CCT in order to do so, a topic which is addressed in **section 5—Outreach**.

Section 6—The link between the Cost Comparison Tool and the Curation Costs Exchange establishes why the CCT forms the nucleus of the CCEx platform and how to benefit from the synergy that arises from that relationship.

Section 7—The future of the Curation Costs Exchange explains succinctly a possible future for the CCEx. Discussions regarding this future are in progress and will not be finalised until the end of the 4C project (January 2015) so no firm conclusions are drawn here.

1 Introduction

In order to make smart investments in digital curation, it is crucial to understand the costs of digital curation, one's own costs at the very least[1].

Amassing empirical data is an obvious starting point if you want to understand a topic, which is one of the approaches the 4C-project has initiated through the part of the Curation Costs Exchange platform[2] (CCEx) that hosts the tool developed for this purpose, the Cost Comparison Tool[3] (CCT). The CCT facilitates the collection of empirical cost data and makes it comparable across organisations, countries, strategies and missions.

We believe that there is a good chance that these comparisons will help practitioners make even smarter investments.

However, the Cost Comparison Tool presents a challenge from the get go because its success depends entirely on the willingness of organisations to share their cost data. In addition, gathering this data can involve a significant effort so there is a further challenge, that of enabling the potential users to understand that the benefits of using the CCT outweigh the initial effort.

We live in an information society where the sharing of knowledge is widely acknowledged to be a key enabler of fast development. However, not only are we asking our stakeholders to share sensitive financial information, we are also asking them to spend time and effort submitting this information in a very structured way to ensure that information from very different sources can become comparable.

This deliverable describes the way that the 4C-project—in collaboration with our stakeholders—has developed a mechanism that makes it possible to compare and analyse in a meaningful way financial information originating from a vast range of different sources and stakeholders.

The deliverable describes a snapshot of the CCT, and even though the tool is a completed and deployed product, we have allowed for its continuous development until the end of the 4C-project (January 2015). We also have plans for the CCEx and CCT to outlive the project in the hands of experienced and knowledgeable organisations who are willing to sustain and continue the development of not only the CCT, but of the whole CCEx platform.

1.1 Definition and scope of the Curation Costs Exchange Framework

The Curation Costs Exchange Framework[4] is a subset of the CCEx platform[5] and represents functionality—embodied by the CCT—which allows stakeholders to submit financial information; to analyse it; and to compare it with information submitted by others.

[1] It is interesting to note that nearly half of those who responded to the relevant questions in the 4C survey relating to costs of curation don't separate the costs of curation from other activities, nearly a third didn't know their annual budget for curation activities, and over a third had never tried to establish the costs (reported in D2.1—Baseline Study of Stakeholder & Stakeholder Initiatives).

[2] http://www.curationexchange.org

[3] http://www.curationexchange.org/compare-costs

[4] "D3.3) Development of the submission template and specification into a workable framework to underpin the function of the Curation Costs Exchange"—4C Description of Work (DoW)

[5] www.curationexchange.org

There is thus a conceptual division between the CCEx platform which deals with how to *inform* stakeholders about the costs and economics of digital curation; and the CCT which deals with how to *exchange* financial information between stakeholders.

Financial information is defined as cost data and the information needed to interpret the cost data:

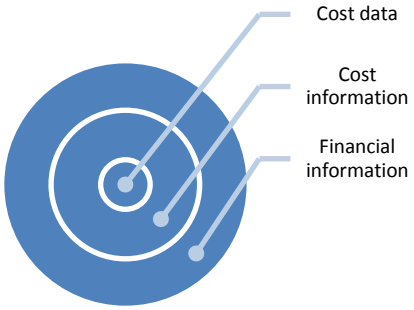


Figure 1—Hierarchy of cost concepts

On the CCEx platform, the CCT can be found in the, "Compare costs"[6] section:

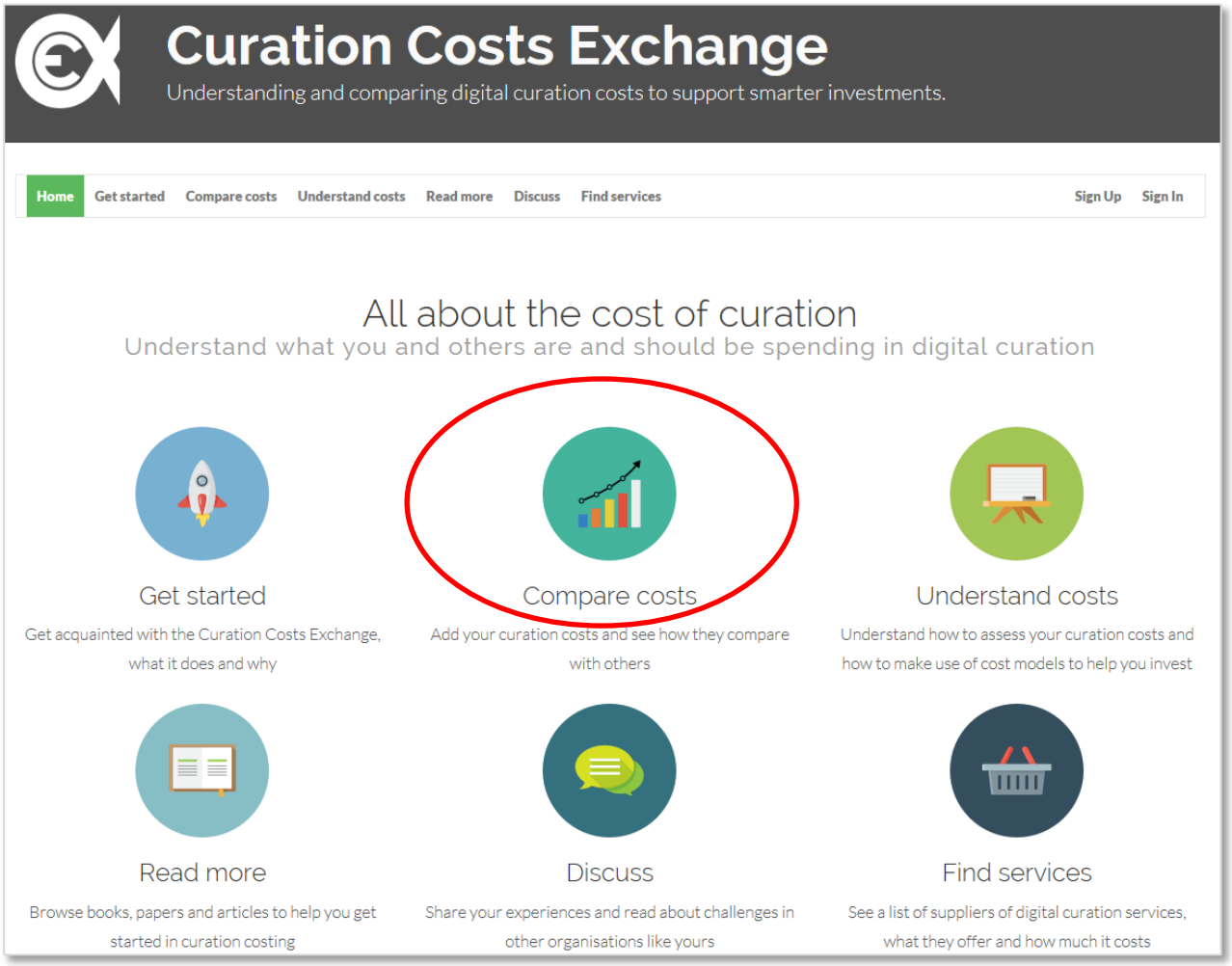


Figure 2—The Cost Comparison Tool

[6] The website is under continuous development so some screen s illustrated in this report may have changed on the live website.

As mentioned earlier, the full deliverable should be considered to be a combination of the written report and the CCT on the CCEx platform.

The definition and scope of this deliverable was described in the Description of Work (DoW):

> *"Building on T2.3 and T3.2 the group will develop a submission template and specification for the Curation Costs Exchange (CCEx). The purpose of this platform is to amass cost and cost related data (e.g. calculation processes, metrics, effort statistics, value calculations, etc.) from stakeholders in order to underpin future activity with empirical knowledge. Experience shows that when anonymity and confidentiality are observed, external stakeholders may even be willing to provide quantitative operational cost data if informed properly about the purpose and use of them. The principal challenge will be to limit the scope of data collection as it is clear that most forms of curation-related information may prove to be an indirect economic determinant: e.g. types of data stored; accession frequency; type of preservation strategy implemented; preservation formats used; access dates, etc. In addition to considering the input to the CCEx, this task will also develop the specification for the output of the system which will dictate how the exchange functions and what is displayed to the user of the platform. Close coordination with T2.6 will be required to implement the framework into a functioning and useable data gathering and dissemination mechanism."*

1.2 Purpose

Being a 'coordination and support action' project, two of the most important missions of the 4C-project are to gather, systematise, analyse and disseminate already existing research on the topic of curation costs, and to rally relevant stakeholders around this material and other outputs created by the project.

The development of the CCEx, including the CCT, addresses this in two ways:

1. By making relevant material and analyses available on the CCEx website;
2. By innovating and incorporating functionality, namely the CCT, which stakeholders can use to share and compare empirical cost data.

These mechanisms—in particular the latter—provide the CCEx with a clearly defined 'unique selling point' that distinguishes it from similar websites, and should attract a broad range of users and sustain user engagement.

1.2.1 The purpose of the Curation Costs Exchange

The overall purpose of the CCEx is to help users understand how to assess the costs of curation at their own organisation. It provides users with financial information[7], which shows them how to adjust plans to get value for money.

Before building the CCEx project partners consulted key stakeholders to identify the main drivers behind the desire to investigate the costs of digital curation (and in the process enable adjustments to business practices to allow cost-effective digital curation). These were[8]:

- to project future costs—85 %

[7] The term financial information covers cost data as well as the cost information necessary to contextualise and interpret the cost data, as seen in Figure 1 above.

[8] http://www.4cproject.eu/d2-1-stakeholders

- to inform decision makers about costs—78 %
- to identify the costs of curating assets—72 %
- to support decision making—63 %
- to assess benefits/value of curation expenses—63 %
- to ensure the efficient use of resources—61 %

These requirements directly or indirectly imply a need to share and compare costs to deliver efficiency to enable informed decision-making, to increase value for money and to reduce expenditure.

Overall, the 4C-project concluded from the consultation that the organisations that have already made a start in costing their curation activities feel that they can do better and are seeking advice on how to fine-tune their costing activities, for example through comparison of cost scenarios. We have also learned that those who are just beginning to explore the costs of curating digital information would appreciate help on getting started and are keen to learn from their peers[9].

The CCEx thus provides a starting point for a range of stakeholders interested in costing curation activities within their organisations. It packages and disseminates information gathered through the project including comparisons of current models, tips for making the case for investment in curation through a better understanding of cost determinants, risks, and benefits and examples of what peer organisations are doing.

Other objectives of the CCEx are:

- to help a range of stakeholders understand which digital curation activities might need to be considered and how they might start to assess these using existing cost models and resources
- to provide digital curators with a platform for communicating with each other in order to clarify costs of curation
- to help develop a cost concept model and a gateway requirement specification[10]
- to enable 4C-project members to understand user behaviour in relation to cost data by assisting them in recognising how to structure the data and identify which parameters are necessary in order to enable comparison of the data

1.2.2 The purpose of the Cost Comparison Tool

The CCEx facilitates the gathering of real financial information from partner organisations and a range of stakeholders. This data is shared (where permitted) using an online submission and analysis mechanism—The CCT. The mechanism aims to capture calculation processes, metrics, effort statistics, and value calculations from stakeholders in order to underpin future activity with empirical knowledge and data sets.

The main purpose for gathering empirical cost data from a wide range of different organisations is that it:

- helps in clarifying the bigger picture of the costs of curation
- constitutes important test data (for future modelling)
- allow us to discover good practice that can be utilised by other organisations

[9] The lessons are learnt not only from the mentioned stakeholder consultation, but also from numerous stakeholder interactions, such as workshops and focus-groups conducted by the 4C-project: http://www.4cproject.eu/community-resources/focus-groups

[10] http://www.4cproject.eu/d3-2-ccm-grs. The two project deliverables (D3.2 and D3.3) have been developed concomitantly and influenced each other.

1.2.2.1 Clarifying the costs of curation, making recommendations and raising awareness

The core mission of the 4C-project is to clarify the costs of curation. Another of the project's missions is to raise awareness about the topic of the costs of curation and make recommendations for the future. However, if we do not know how organisations deal with their costs today, we cannot hope to understand their processes and consequently we cannot hope to make appropriate recommendations.

Understanding the costs of one's curation can help improve decision making on many levels—operational, tactical, strategic and political. Improving future practices requires wide ranging knowledge of the situation 'as is' (from as many different 'situations' as possible) and some mechanism to discover, record and repurpose good practices in other organisations.

1.2.2.2 Gathering test data

Test data for development of internal deliverables

The cost data gathered to date has helped the 4C-project develop the Cost Concept Model (deliverable D3.2) and the CCEx (deliverable D2.8). Both deliverables require knowledge of how financial information can be best understood, structured and communicated. It is vital—as in any other development process—to couple the theoretical approach with empirical data.

Test data for attracting users to the CCEx

Populating the CCEx with as much cost data as possible will make it more attractive to the user. As data volumes increase, the outputs of the CCT will become increasingly accurate, diverse and relevant for additional stakeholder groups. This in turn allows users to start comparing costs to make smarter investments.

Test data for the development of cost models

One of the biggest flaws of the existing cost models is their lack of accuracy. Another is that their scope is too narrow; most models address one organisation type only. It is more than likely that one of the reasons for this is that models have been developed partly or wholly without using empirical data. The CCEx is different in both respects: it enables cost model developers to test and perfect cost models and to account for different organisations and scenarios.

1.2.2.3 Good practice

The CCEx seeks to be a discussion platform for the community. By providing a communications hub alongside the CCT it is hoped that we will be able to encourage the cross community sharing of good practice by all stakeholders.

1.3 Terminology

4C terminology[11]—and to the extent possible OAIS terminology—are used as a baseline in this report where appropriate.

[11] http://www.4cproject.eu/community-resources/glossary/full-glossary

1.4 Prerequisite of the comparison of costs

The CCT is a crowd-sourced database for the sharing of cost data. As such its success depends entirely on the willingness of organisations to submit their cost data. In order to achieve a critical mass of data (which will to encourage greater use of the CCEx) the 4C-project has undertaken the collection of cost data from 4C consortium partners as well as from external stakeholders. Efforts are on-going and we aim to have 25 cost data sets by the end of the project as discussed in section 5.1.1, Measures of Success.

1.5 Audience and benefits

The stakeholder groups of the 4C-project[12] are the same as but not necessarily limited to the target audience for the CCEx, which caters for 11 key stakeholder groups:

1. Research funders
2. Big data science
3. Digital preservation vendors
4. Government agencies
5. Publisher or content producers
6. Data intensive industries
7. Memory institutions or content holders
8. Small or medium enterprises
9. Universities
10. Researchers
11. Other

In order to better understand what each stakeholder group would look for in the CCEx, use cases have been generated. These examinations of user behaviours and needs target not only the users of the CCT, but of also the whole CCEx. The use cases also provided information on the benefits of using the CCEx. These benefits emerge from the table below, and also appear in a distilled version on the CCEx platform[13].

[12] http://www.4cproject.eu/d2-1-stakeholders

[13] http://www.curationexchange.org/get-started#benefits

Stakeholder Group	Exemplar Role	Information sought in CCEx	Reasons for using CCEx	Initial CCEx information provision	Main benefits	Information to share	Anonymised?
Research funder	Funding programme manager; provide guidelines to peer reviewer of grants	I would expect to find out what kinds of activity costs might be eligible to be covered in-project (what and how much time is "reasonable").	I want to ensure that our guidance helps applicants to understand eligible costs. I want to make sure that the research funds are used in the most cost-effective way and yield impact (Return on Investment).	Distillation of UKDA14 cost tool (activities and guidance on when costs are most effective—in-project or at Ingest to an archive); Overview of KRDS15 activity models as reference to activities to consider; Link to DCC16 funders' policy overviews.	The Cost Comparison Tool enables the identification of activity costs which might be covered 'in-project' e.g. what/how much time is "reasonable," to ensure that funder guidance helps applicants to understand eligible costs, and that research funds yield the best return on investment.	I'd be willing to share our policies on including costs and provide pointers to guidance.	No
Big data science	Research Infrastructure (RI) provider	I want to understand how curation costs should be included in my operational budget and to find and assess suitable third curation party services we might need to procure.	I want to be able to ensure that my RI facility operates efficiently and to be able to plan for sustainability.	Comparison with other data stewards (data centres, national libraries).	The Cost Comparison Tool enables the identification of relevant costs in terms of domains, scale and activities, to assist with effective budgeting for new projects and making reasonable estimates from the outset.	I'd be willing to provide costs relating to specific services offered by our RI facility.	Possibly
			I want to find the right service provider for my particular needs and be able to understand their charging models.	Links to preservation and curation vendors and services.		I'd be willing to share feedback on the level of satisfaction with the procured service or product.	Possibly

[14] UK Data Archive

[15] Keeping Research Data Safe

[16] Digital Curation Centre

Stakeholder Group	Exemplar Role	Information sought in CCEx	Reasons for using CCEx	Initial CCEx information provision	Main benefits	Information to share	Anonymised?
Digital preservation vendor	Preservation service provider	I would expect to see real cost data related to operating a digital preservation service (either as a third party service or in-house solutions)	I'd like to compare my expenditure with others and learn how others are spending their budget. I want to see how I compare with competitors and peers and understand the reasons for significant disparities in costs, so I can implement processes to reduce my high cost activities	Access to real cost data.	The Cost Comparison Tool shows real cost data related to operating a digital preservation service (either as a third party service or in-house solutions) and enables comparisons of expenditure with that of competitors and peers, providing a greater understanding of the reasons for significant disparities in costs, and enabling the implementation of possible remedial activities.	I'd be willing to share my own cost data to get a more accurate result in the comparison.	Yes
Government agency	Preservation manager of a National Library	I would be most interested in browsing through cost model descriptions to compare my expenditure with others. I'd like to know which cost model might best meet my organisation's needs. I'd like to see how other National Libraries define curation related risks and benefits.	I'd like to support proper planning of a new content stream being taken into the archive.	Overviews of current cost models.	The CCEx provides cost model descriptions that will help to inform which cost models to use. It will enable comparison of expenditure between government organisations and assist with planning and budgeting.	I would be willing to share information on cost models we employ.	No
			I'd like to compare the costs of preserving different types of collection content (e.g., an e-journal collection).	Suggestions on which models are suitable for specific types or organisations and/or content.		I would be willing to share feedback on suitability of cost model for my organisation /collection type.	No
			I want to find out if other National Libraries are spending their budgets more efficiently (e.g., preserve larger data holdings with fewer resources) so I can undertake an informed review of the efficiency of our own workflows.	Comparison with other data stewards (data centres, national libraries).		I'd be willing to contribute our own curation cost data to help develop more accurate comparisons between National Libraries.	Possibly

Stakeholder Group	Exemplar Role	Information sought in CCEx	Reasons for using CCEx	Initial CCEx information provision	Main benefits	Information to share	Anonymised?
Publisher or content producer	Repository manager	I would expect to find out how I might be able to reduce my curation costs and increase efficiency in service delivery.	I'd like to compare my costs with similar repositories.	Access to real cost data.	The Cost Comparison Tool enables comparison of cost structures with those of other companies, highlighting areas of potential cost-reductions in day to day activities.	I would be willing to input my operational and/or cost data.	Yes
Data intensive industry	Financial department of large broadcasting company	I want to find out what other large national organisations are spending to preserve access to their digital collections.	I'd like my company to be able to guarantee access to our data for as long as it is required in a cost-effective, risk free (or at least risk minimised) fashion. If possible, I'd like to see a return on our investment.	Access to cost comparison tool and short case studies.	The Cost Comparison Tool will show what large national organisations are spending to preserve access to collections so that access may be guaranteed to data for as long as it is required, while reducing risk in a cost-effective way.	I would be willing to share information on our curation services and the costs or operation.	Possibly
Memory institution or content holder	Community archive manager	I would expect to see general information that would help me to manage limited budget resources.	I've been tasked with doing some form of planning or analysis exercise which involves examining the costs / benefits / business case for future digital curation activity and I don't know how to start.	Getting started guidance. Overview of cost models, risks and benefits. Access to cost comparison tool and short case studies.	The Cost Comparison Tool offers controlled contact and comparisons between large and small institutions with a view to assisting in the management of limited budget resources, creating business cases for future activity and benchmarking—with accuracy indicators to manage the different scales of operation.	I'd be willing to cooperate with and share advice and guidance with other smaller institutions.	Possibly
			I'd like to know what larger institutions are spending in certain areas, but I'm more interested in what other smaller institutions are spending. I'd like to find information that will allow me to benchmark our approach.			I have some historical costs data and/or emerging estimates of costs that I'd be willing to share to help other smaller institutions.	Possibly

Stakeholder Group	Exemplar Role	Information sought in CCEx	Reasons for using CCEx	Initial CCEx information provision	Main benefits	Information to share	Anonymised?
Small or medium enterprise	DP practitioner that is relatively new to the field	I would like to browse through cost models to find out what is out there and is potentially useable by my organization.	I'd like to be able to propose a viable approach to senior management at my organisation outlining a feasible approach to implement (make the case).	Access to a list of risks and benefits relating to digital curation. Advice on how to develop and integrate cost modelling into organisational budgeting and accounting methods.	The CCEx provides an assessment of different available cost models in order to identify which is the most suitable for a particular type of organisation.		
University	Research administrators(RA), IT managers, library staff	RA—I would like to know what to budget into new grant proposals to cover curation costs during the active phase of research. IT and library staff—we would like to make sure we can afford to provide and sustain curation services and storage in line with researchers' needs.	I would like to be able to cost curation activities into new proposals. I would like to be able to budget for sustainable service provision.	Access to activities that should be costed into proposals. Access to cost information for peer organisations.	Institutions would be better able to plan for and sustain curation costs across the research lifecycle.	RAs—I would provide details on financial systems we use within the institutors. IT and library—we would share our costs so we can see how we compare with our peers.	Possibly
Researcher	Principal Investigator	I would expect to find out what kinds of activity costs might be eligible to be covered in-project (what and how much time is "reasonable"). I want to see costs relevant to me in terms of domains, scale and activities.	I want to be make a good case for justifying additional resources in my application, so I can ensure I can get funded and I meet funders' expectations about Research Data Management and sharing.	Access to guidance on curation activities to consider when developing new project proposals.	To compare with other, similar projects; to see where I could bring my costs down; to comply with funders' expectations, both regarding data management, preservation and expenses	I'd be willing to share information about the kinds of curation activities that I'm planning to undertake. I might be willing to share my data management plan (DMP).	Possibly

Table 1—Use cases for the CCEx

2 Description of the Cost Comparison Tool

As previously mentioned, the CCT is one of the core services accessed through the CCEx. As such it is prominent on the opening page of the CCEx and can be accessed via the 'Compare costs' icon[17]:

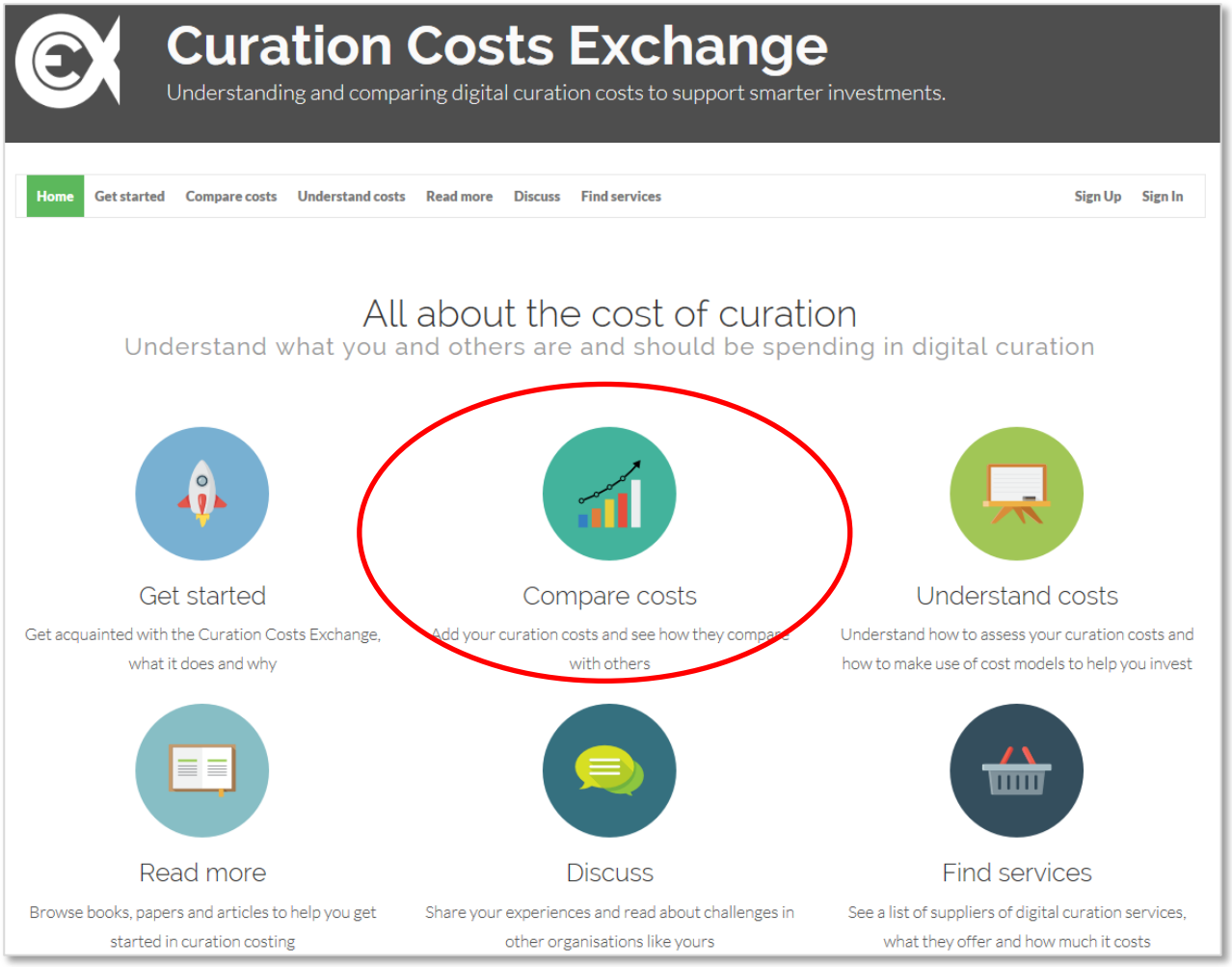


Figure 3—The Cost Comparison Tool on the Curation Costs Exchange front page

In the CCT, users submit and compare cost data via a step-by-step process as described in the following sections.

2.1 User input: Add cost information

Figure 4—User input: Add cost information

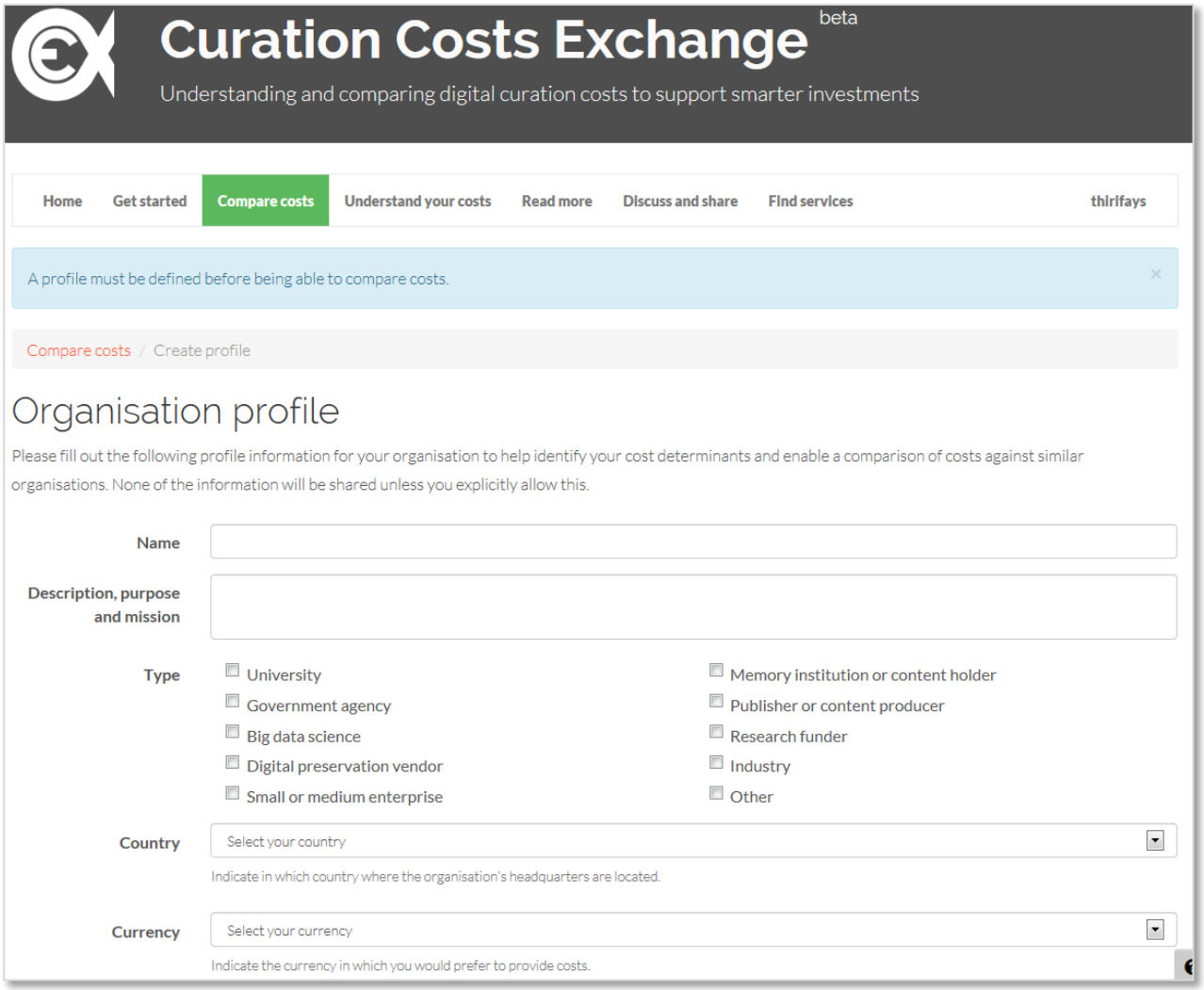
[17] Because the 4C-project has focused on coherence and synergy between its deliverables, there is inevitably a high degree of overlap. This interrelationship is addressed elsewhere in this report, cf. Section 6 The link between the Cost Comparison Tool and the C.

Cost information is more than just numbers. The fact that an organisation may be spending €200,000 a year on curation related activities is not enlightening unless you have more background information about the organisation and its digital assets (for example organisation size, type, data volume, asset types, country of residence and so on). We have defined cost information in this context as information that contextualises and explains the cost data that users submit.

The submission of cost information is divided into 2 steps:

1. Organisation profile: Information about the organisation
2. Cost data profile: Information about the scope of the cost data and the nature of the digital assets[18]

2.1.1 Organisation profile: Information about the organisation



[18] This is what the cost data covers – a whole organisation, a department, a project, a collection—as well as information on the digital material itself.

Figure 5—Organisation profile: Information about the organisation

The user is invited to submit the following information († indicates optional information[19]):

- **Name**—to identify the organisation
- **Description, purpose, mission**†—to accommodate stakeholders' desire to nuance the cost data they submit
- **Type**—to enable comparisons within a specific stakeholder group
- **Country**—to enable comparisons in a specific country; and to take in to account disparate salary levels
- **Currency**—to enable conversion to the default underlying currency that the comparisons are based upon (€)

2.1.2 Cost data profile: Information about the scope of the cost data and the nature of the digital assets

The first part of the cost data profile covers information about the scope of the data, in other words the defining extent of the data set—for example a project—and period the information covers.

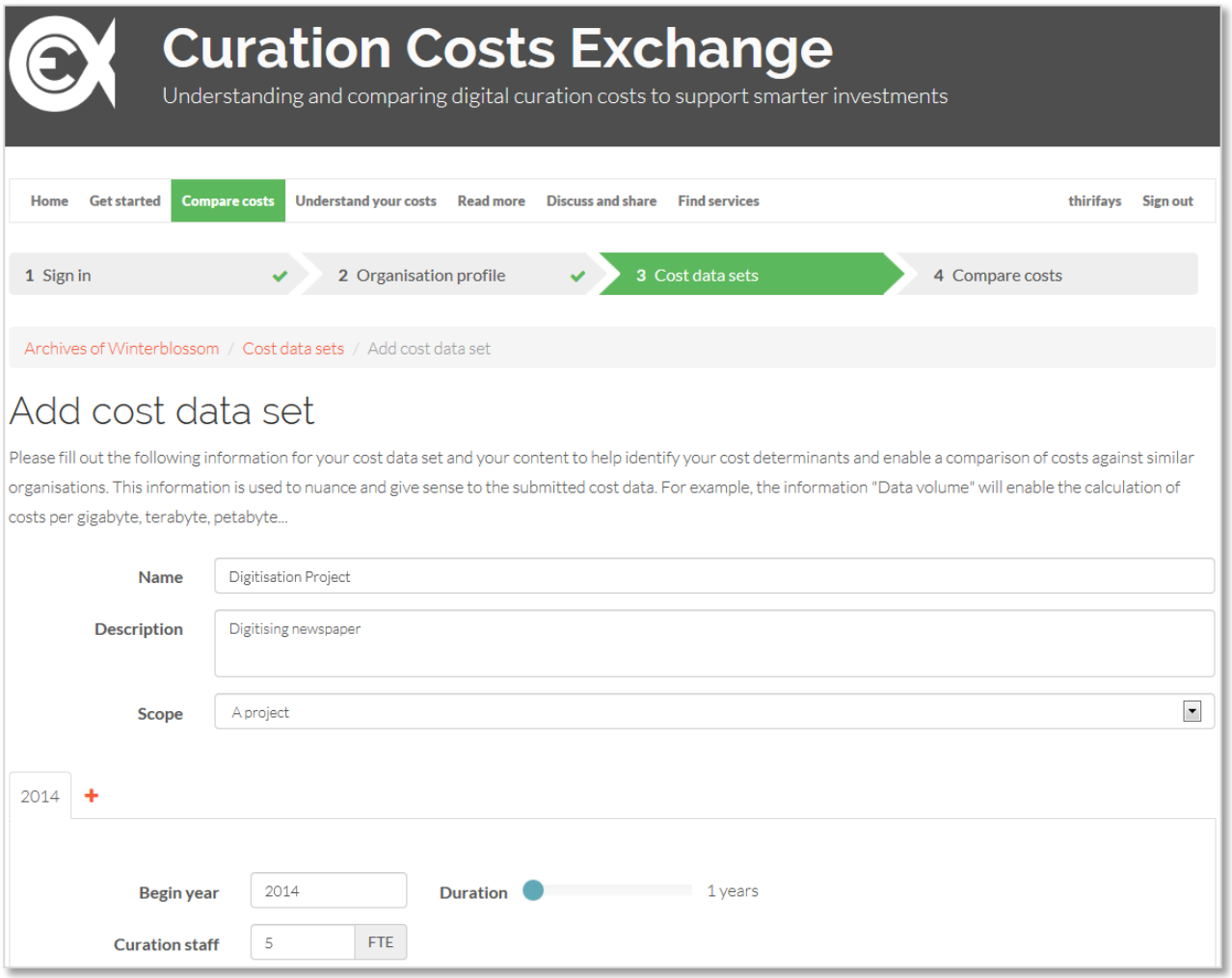


Figure 6—Cost data profile #1: Information about the scope of the cost data

[19] A convention that is used throughout the rest of this document

Data collected includes:

- **Name**—to identify the cost data set
- **Description**†—to accommodate stakeholders' desire to nuance the cost data they submit
- **Scope**—to define what the cost data covers.
  Options here are: The whole organisation, A department, A project, A collection, Other
- **Begin year** (**Duration**)—to place the cost data in time and to determine if it is current
- **Curation staff** (**FTE**[20])—to enable comparisons concerning staff size

The second part of the cost data profile covers information about the nature of the digital assets, the volume and the types of the digital material being curated.

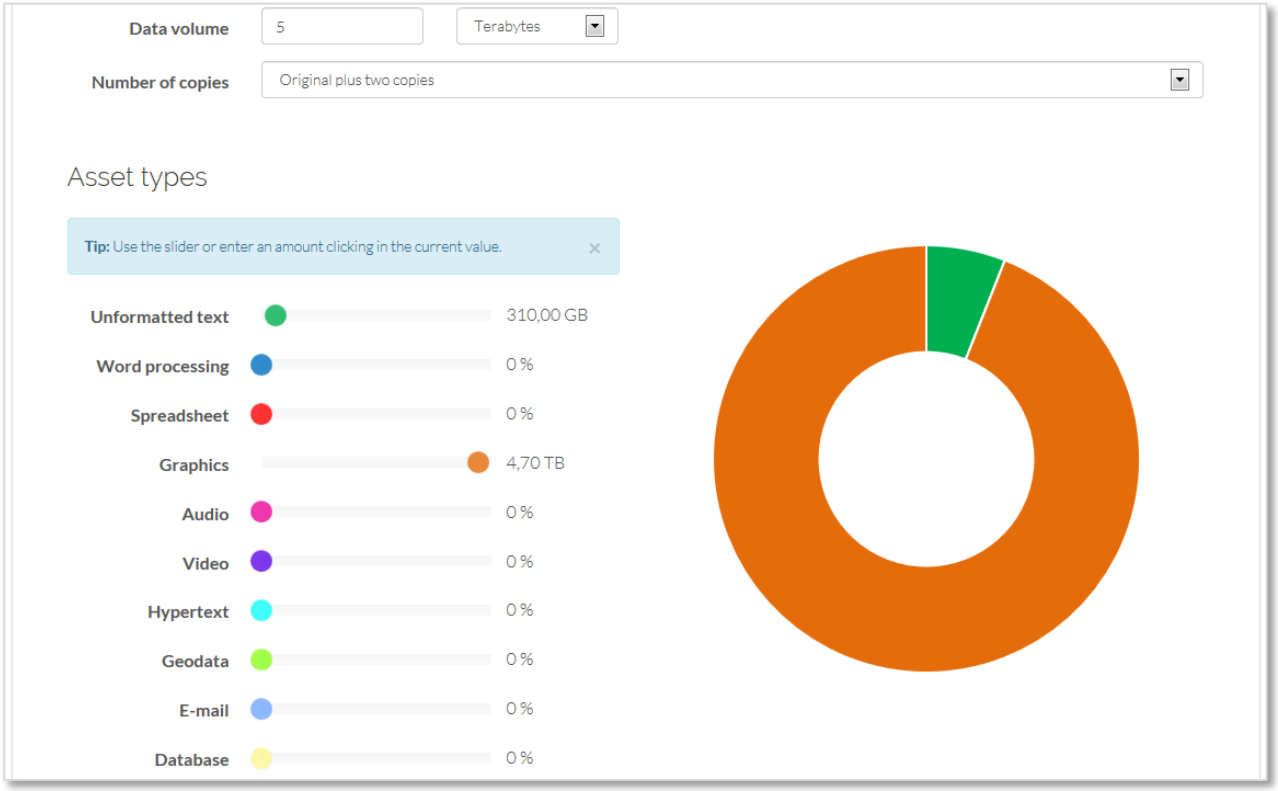


Figure 7—Cost data profile #2: Information about the nature of the digital assets

Initially users set the following:

- **Data volume**—to enable comparisons regarding volume
- **Number of copies**—to enable comparisons between different strategies. This also affects data volume
- **Asset types**—to enable comparisons between different formats

As data is added users can allocate data across types (see Table 2 below for information regarding the types—the 'Variables/options')).

[20] FTE—Full-time equivalent. A representation of the number of full time staff allocated to a task that takes into account the fact that some staff may only be partially engaged in the task. For example, three members of staff engaged on a task each contributing ⅓ of their time to the task equates to one FTE.

These pieces of information serve both as filters and cost determinants. As filters, the information is used when the cost data is displayed at the end of the process. An example of such a filter is data volume which enables an organisation to compare its cost data with that of an organisation curating approximately the same volume of data.

As cost determinants, the information is used to contextualise the cost data. For example high data volume *could* explain high costs. In order to understand which cost determinants should be included in the Cost Comparison Tool, three methods were employed:

- Stakeholders were consulted to establish which pieces of information are necessary in order to understand cost data
- We drew upon expert knowledge within the 4C-project many of whom are digital curation cost experts who have played an active role in the development of cost models
- We performed analyses of the impact and interplay of the various cost determinants

The initial gross list held more cost determinants than the CCT incorporates today. However, asking stakeholders for too many pieces of information would discourage users from using the tool so the list of determinants has been prioritised.

The gross list included (elements in **bold** have been retained):

Cost information element	Variables/options
Costing behaviour	How do stakeholders account for their expenses, and which kinds of negative and positive effects do these behaviours hold?
Asset quality	Free text
Asset types	Unformatted text, word processing, spreadsheet, graphics, audio, video, hypertext, geo-data, e-mail, database
Asset volume	Gigabyte (GB) intervals
Core business activity	Is digital curation is a core activity? Y/N
Country	Select from a list of countries
Curation staff	The number of people working with the digital assets defined in the scope
Indirect Cost Drivers[21] (ICD's)?	Is an organisation spending its money on acquiring a good reputation? On risk management? On preserving authentic records? On innovation?
Main funding source	Private/public/Other
Motivation for curation	Legal, business, research, profit
Number of copies	1/2/3/4/more
Organisation size	Selection of range of number of employees
Outsourcing of activity	Yes/No/NA
Preservation strategy	Emulation / migration/ not applicable
Quality of repository	Free text
Scope	What does the cost data cover (organisation, department, collection, project)?
Stakeholder group	Select from a list of stakeholder groups
Stakeholder mission	Optional free text entry (also covers organisation mission)
Time span	Long-term/medium-term/short-term (to be defined in years)

[21] http://www.curationexchange.org/make-the-case/16-indirect-cost-drivers

Cost information element	Variables/options
Trusted digital repository (TDR)	Yes/No/NA

Table 2—Cost information elements (filters and cost determinants)

2.2 User input: Add cost data

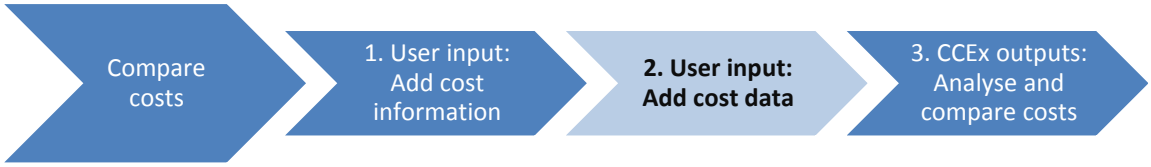


Figure 8—User input: Add cost data

Once the user has submitted the cost information, it is time to submit the associated cost data and to normalise[22] it. In the CCT cost data covering a specific scope (such as an organisation, a department, a collection or a project) is known as a 'cost data set'. Each cost data set is divided into 'cost units' that represent the components in which the organisation has separated their costs:

Cost data set (e.g. the costs of a migration project)	
Cost unit name	**Expenditure**
Cost unit #1, (e.g. the costs of Project Management)	X €
Cost unit #2 (e.g. the costs of the Migration Process itself)	X €
Cost unit #3, (e.g. the costs of Quality Assurance)	X €
Total	**XXX €**

Table 3—Typical source cost data

However, the ways that organisations account for their expenses varies and, in order to make the expenses of different organisations comparable, it is necessary to express them in a uniform way. This cost data normalisation consists of the aggregation of submitted cost data into categories predefined by the 4C-project. Although this process allows data to be compared in a consistent fashion there is inevitably some loss of accuracy and detail. However, we feel that this loss in fidelity is minor when set against the other processes it enables us to do.

The normalisation process is divided into two steps to cater for different costing practices across organisations. In the CCT, it is mandatory to perform at least one of the two normalisation processes, and optional, but recommended, to perform both. The processes are:

1. Normalisation of costs by activity: Pre-Ingest, Ingest, Archival storage and Access
2. Normalisation of costs by resource type: Capital, labour and overhead

[22] The normalisation process consists of aggregating different cost data into cost categories predefined by the 4C-project. This allows us to compare costs irrespective of the way the data was originally organised by the user's organisation.

An explanation of how these categories are defined and were decided upon is covered in section 3—Rationale and mechanisms.

2.2.1 Normalisation of costs by activity: Pre-Ingest, Ingest, Archival storage and Access

This step allows for the submission and normalisation of an organisation's curation cost data into activity-based categories—past costs are related to 'standard' activities.

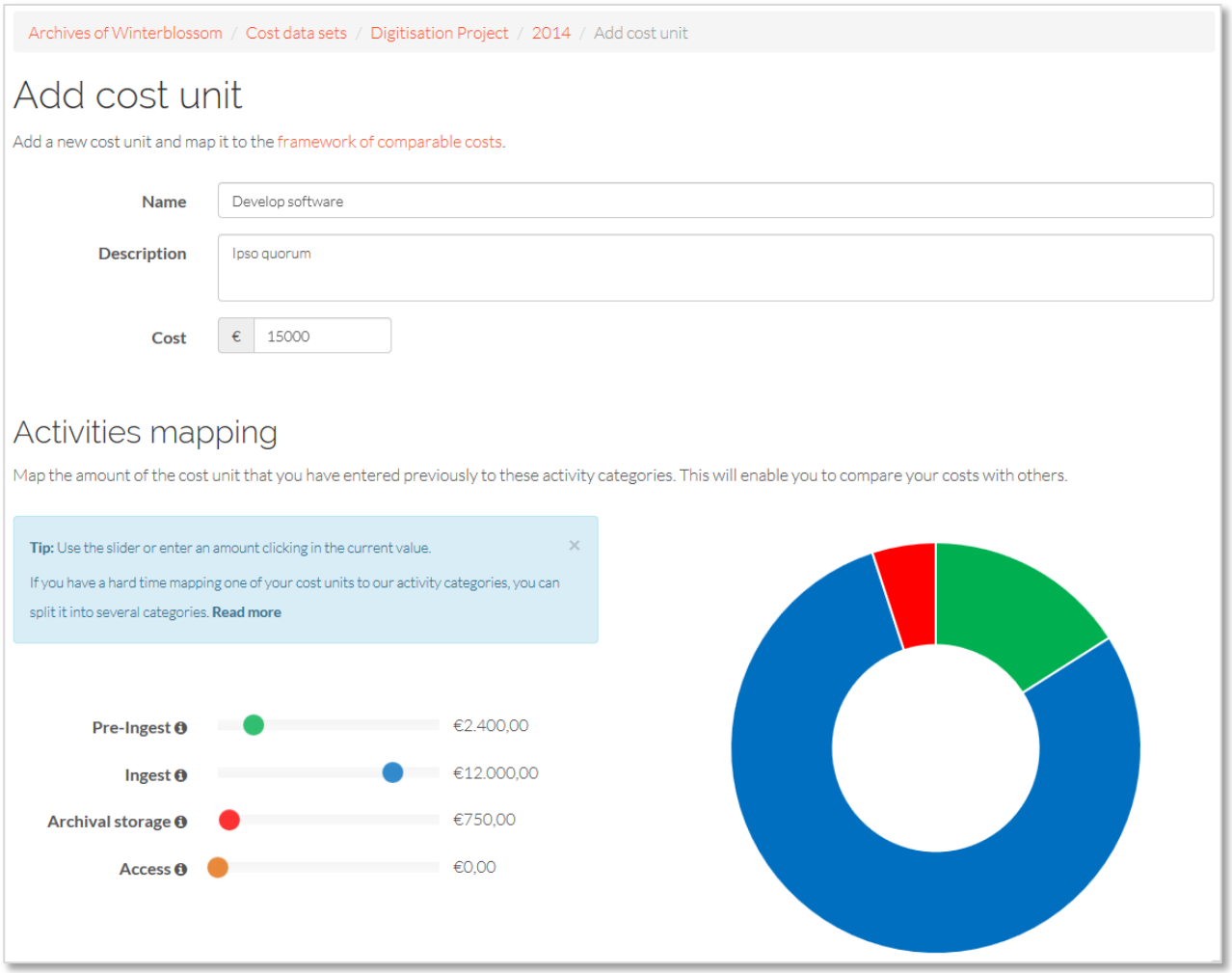


Figure 9—Normalisation of cost data into activity-based categories

Figure 9 above illustrates a hypothetical digitisation project which costs €50,000. The financial department has broken this cost into more detailed figures (named 'cost units' in the template) as follows: Buy hardware (€10.000); Develop software (€15.000); Scanning (€10.000); Project management (€10.000); Quality assurance (€5.000). When examined more closely it becomes apparent that some of these costs should be allocated across more than one 'standard' activities. For instance the cost unit in Figure 9 above—Develop software (€15.000)—spreads across Pre-ingest, Ingest and Archival software.

As cost units are submitted one by one the normalisation process allows the user to allocate all or part of the cost to one or more of the activity categories proposed by the CCT (in our example Pre-Ingest, Ingest, Archival Storage, Access). The cost of 'Develop software (€15.000)' is in this example allocated to Pre-Ingest (€2.400), Ingest (€12.000) and Archival Storage (€750).

The process is:

- **Add cost unit**—Add the first cost unit (this is repeated as many times as you have cost units)

- **Name**—Name of cost unit, for example 'Develop software'
- **Description**†—Description of the cost unit, for example name of scanner
- **Cost**—The actual cost of the cost unit

- **Activities mapping**—Normalise or 'map' the amount of the cost unit to one or several of the proposed activity categories:
  - **Pre-Ingest**
  - **Ingest**
  - **Archival storage**
  - **Access**

An explanation of how these categories are defined and were decided upon is covered in section 3—Rationale and mechanisms.

The sliders allow for the allocation of the costs of one cost unit into one or more activities. If, for example, the 'Develop software' activity covered the development of programs for both Ingest and Access activities, the costs of the 'Develop software' cost unit could be distributed evenly (50%) (or unevenly as appropriate—e.g. 25%-75%) between Ingest and Access.

2.2.2 Normalisation of costs by resource type: Capital, labour and overhead

This step is designed to facilitate the submission and the normalisation of an organisation's curation cost data into categories based on resource type. Past costs are expressed in terms of capital costs ("Purchases"), labour costs ("Staff") and indirect costs ("Overhead").

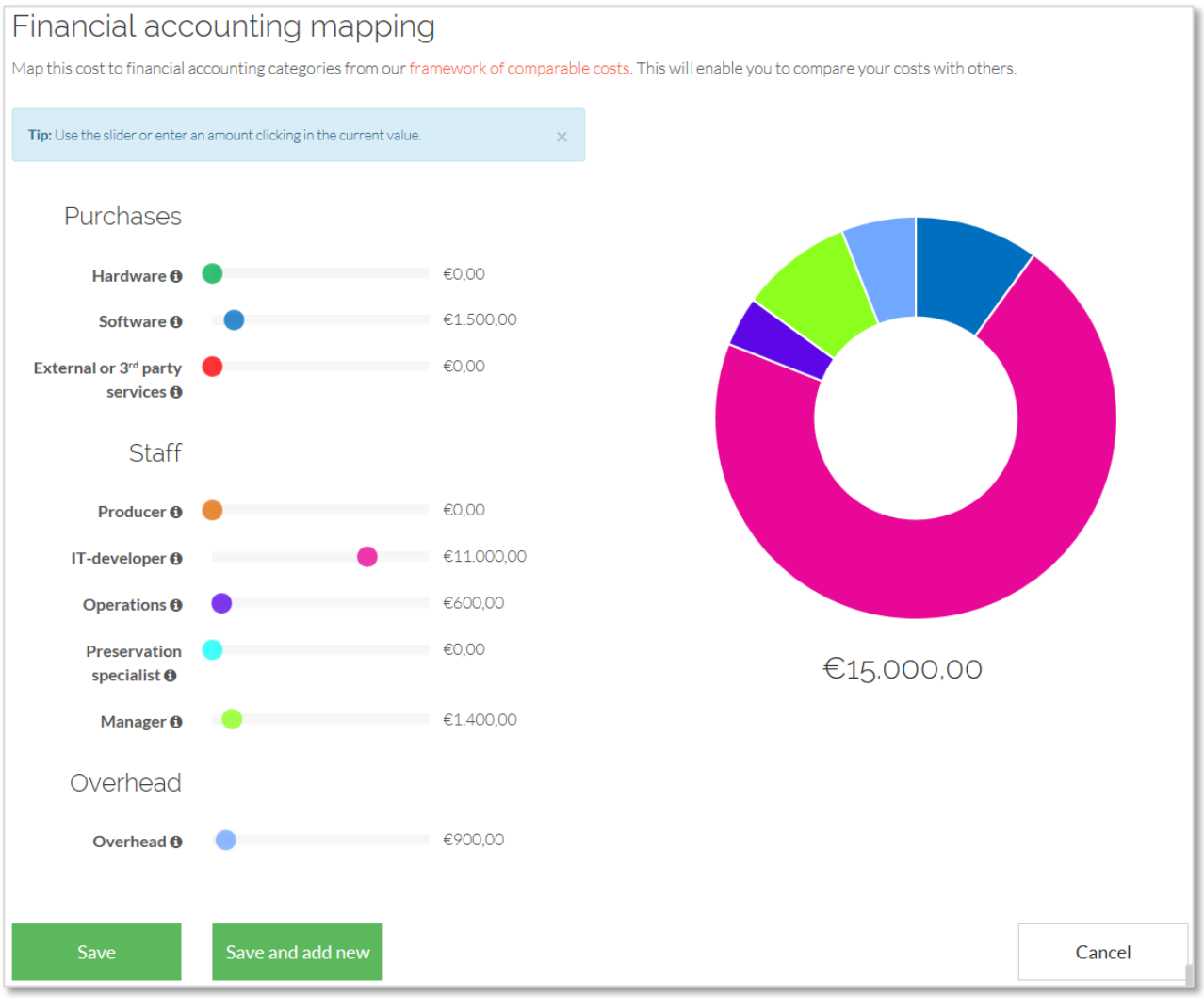


Figure 10—Normalisation of cost data into resource type categories

This normalisation process is the similar to that for the activities. In this case the cost relating to each cost unit is allocated to the proposed sub-categories of "Purchases" and "Staff". Any indirect costs can be allocated to the category "Overhead".

An explanation of how these categories are defined and were decided upon is provided in section3—Rationale and mechanisms.

2.3 CCEx outputs: Analyse and compare costs

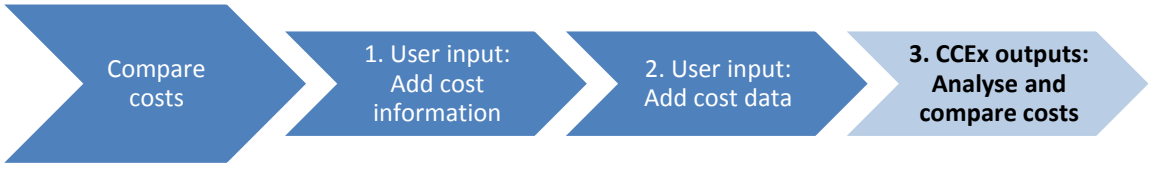


Figure 11—CCEx outputs: Analyse and compare costs

2.3.1 Preliminary results and making figures public

When the data input steps have been completed and the costs have been normalised users are presented with a preliminary result along with the option to 'finalise' their cost data (see below). In this context finalise makes the information available for purposes of assessment and comparisons. Users will also be

shown to what degree the normalisations are complete—the columns 'Map to activities' and 'Map to resource type'[23] display a percentage indicating the degree of completeness of the normalisations.

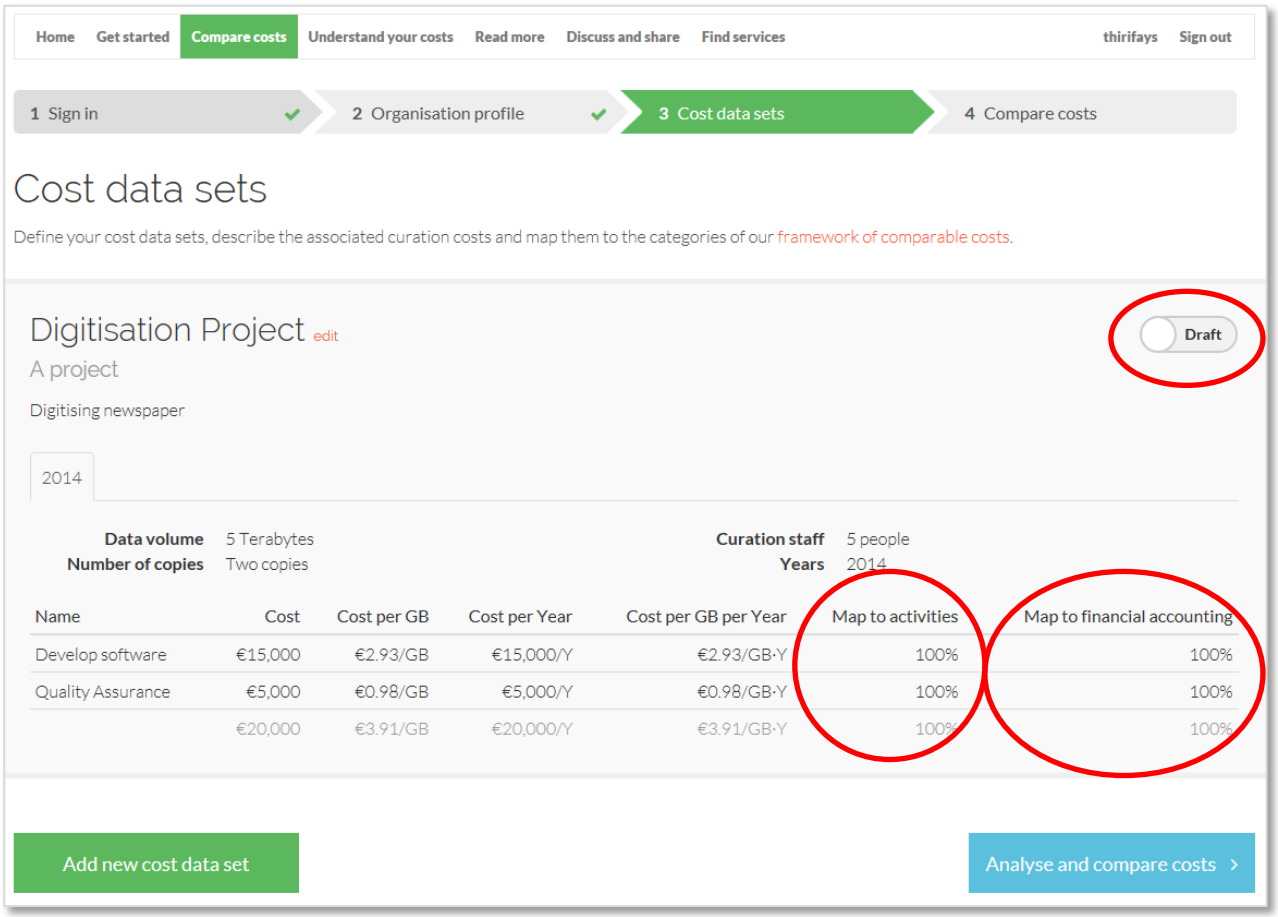


Figure 12—Preliminary results and making costs public

Clicking the 'Draft' button will publicise the cost data, incorporating it in the analyses and comparisons described in the following section.

Figure 13—Draft to complete—private to public

2.3.2 Results

Once the user has submitted the cost information and normalised the cost data, the CCT automatically analyses the data and displays a series of results that can be used for self-assessment, group comparisons, and individual comparisons. Users can access these result types by clicking on the tabs "My costs"; "Global comparison"; and "Peer comparison":

[23] Note that 'resource type' corresponds to 'financial accounting' throughout this report on the screenshots. When writing this report, this change had not yet been implemented on the website.

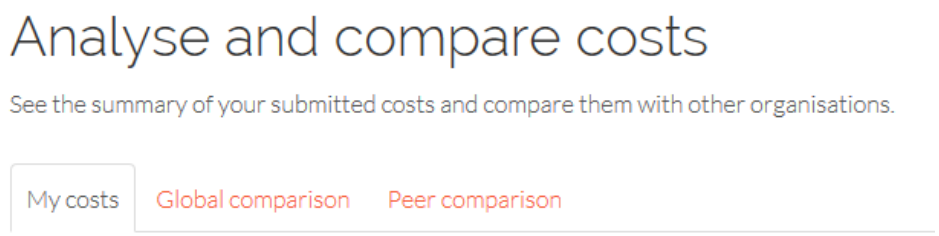


Figure 14—Three types of results: Self-assessment, group comparisons and individual comparisons

2.3.2.1 My costs: Self-assessment

The purpose of the self-assessment is to enable analysis of one's own expenses and to help reflect on the costs of curation. It is also possible to do in-house comparisons if the costs of, for example, several projects have been submitted to the CCT.

The first option encountered under this tab is to choose which costs to display (assuming that the same organisation has submitted several cost data sets). In the example below we have two fictional sets, the "Digitisation Project" and the "Archives of Winterblossom".

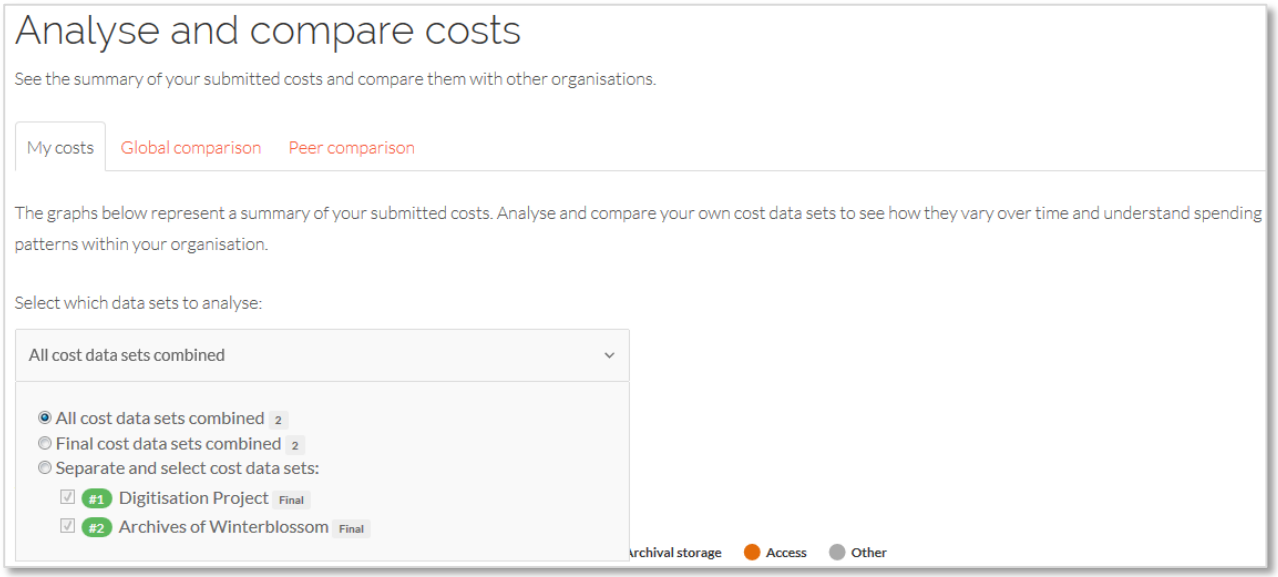


Figure 15—Overview of submitted costs

When the costs to be displayed have been selected, two graphs can be consulted—one reflecting the activity-based approach and one reflecting the resource type approach:

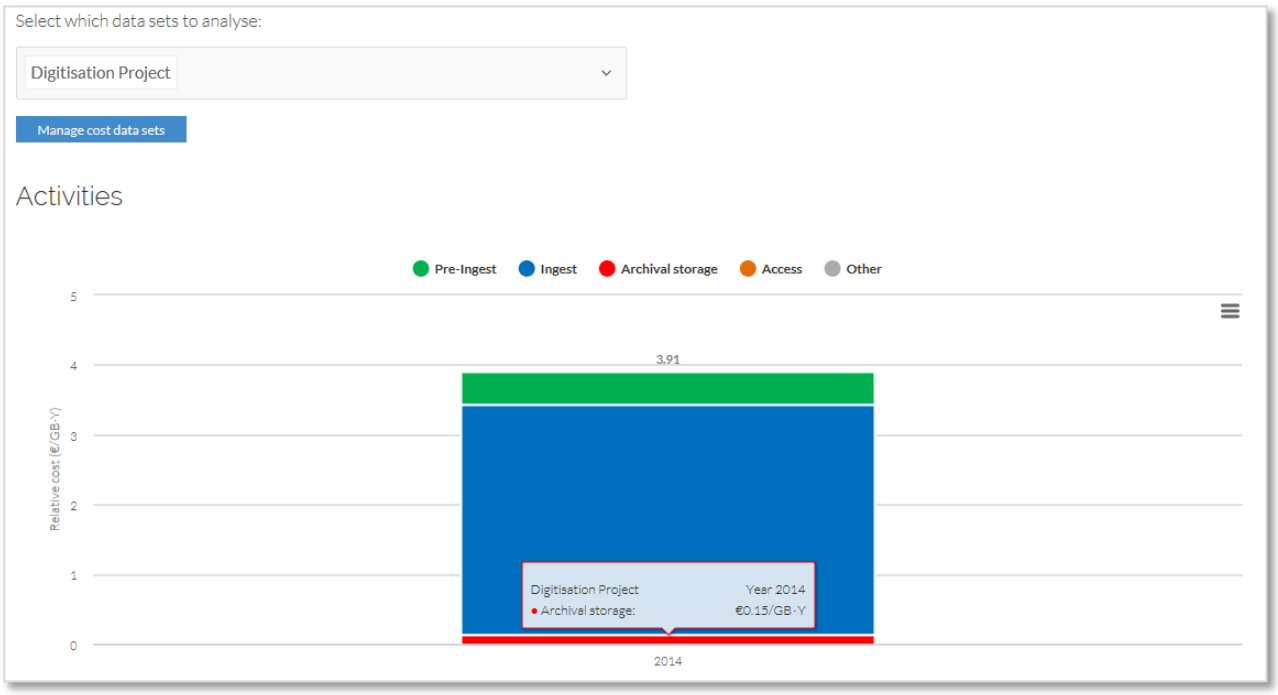


Figure 16—Graph displaying self-assessment regarding activity-based costs

As the name implies the "Activities" graph provides an overview of the aggregated costs according to activity-based costing principles. They are expressed as the relative costs (€ per GB per year) of each activity category. It is possible to hover over each section of the bar or use the key to retrieve accurate information, as shown in Figure 16 above. The number at the top of the bars shows the total expenditure per year. Data can also be exported by clicking the icon in the top right-hand corner of the graph.

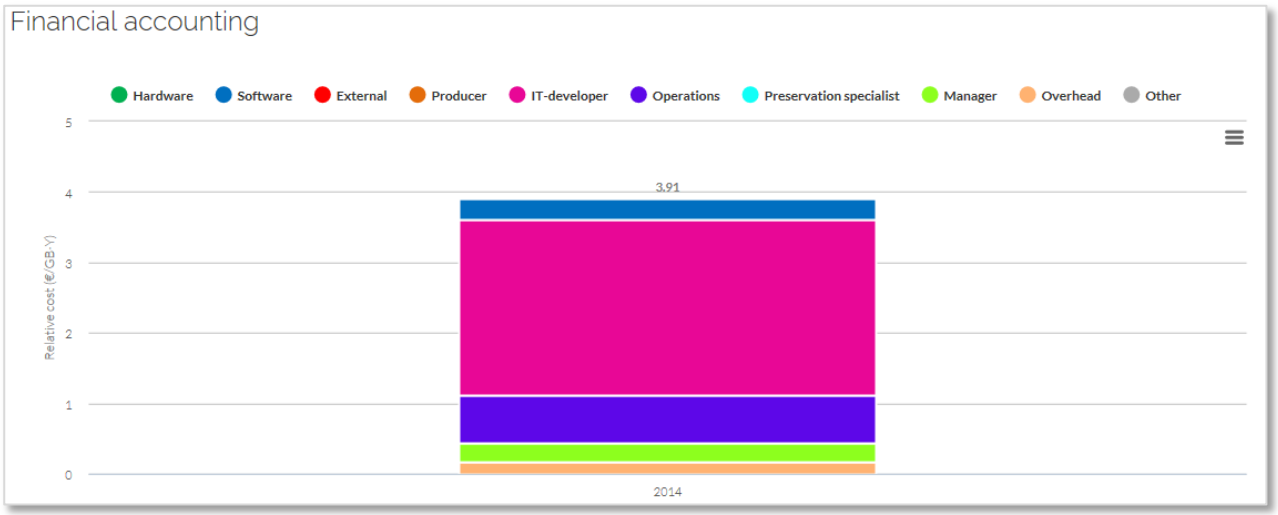


Figure 17—Graph displaying self-assessment regarding costs expressed as resource types

The "Financial Accounting" graph provides an overview of the aggregated costs according to costing by resource type principles. The graph can be manipulated in much the same way as the one described above.

2.3.2.2 Global comparison: Group comparisons

The group comparisons compare cost data of one organisation to a global average. This comparison can be further refined by using filters, see Figure 18 below. The purpose of the group comparisons is to

enable identification of similarities and discrepancies in curation costs between organisations so that stakeholders can detect where there is potential for streamlining and improving outcomes.

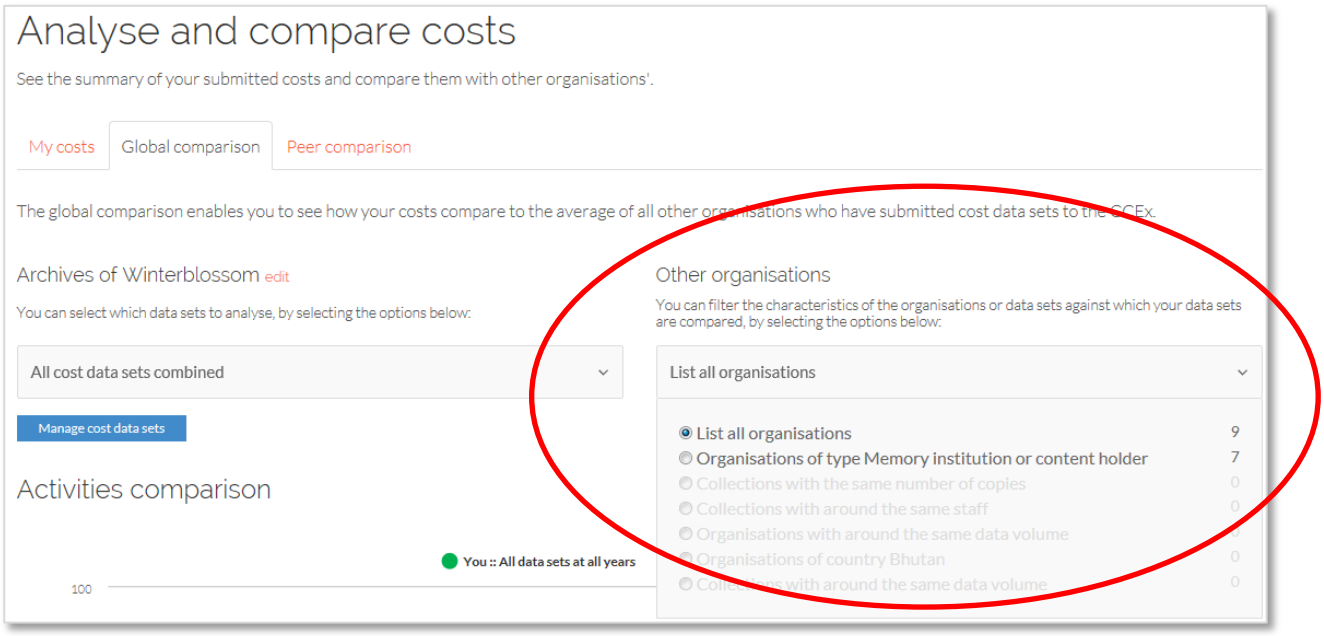


Figure 18—Selection of filters to be included in the global comparison

The filters match the ones already presented in Table 2—Cost information elements (filters and cost determinants), except for the ones that are not quantifiable (such as stakeholder mission).

When the costs to be displayed have been selected and the filters have been applied, two graphs can be consulted—once again one reflecting the activity-based approach and one reflecting the costing by resource types approach.

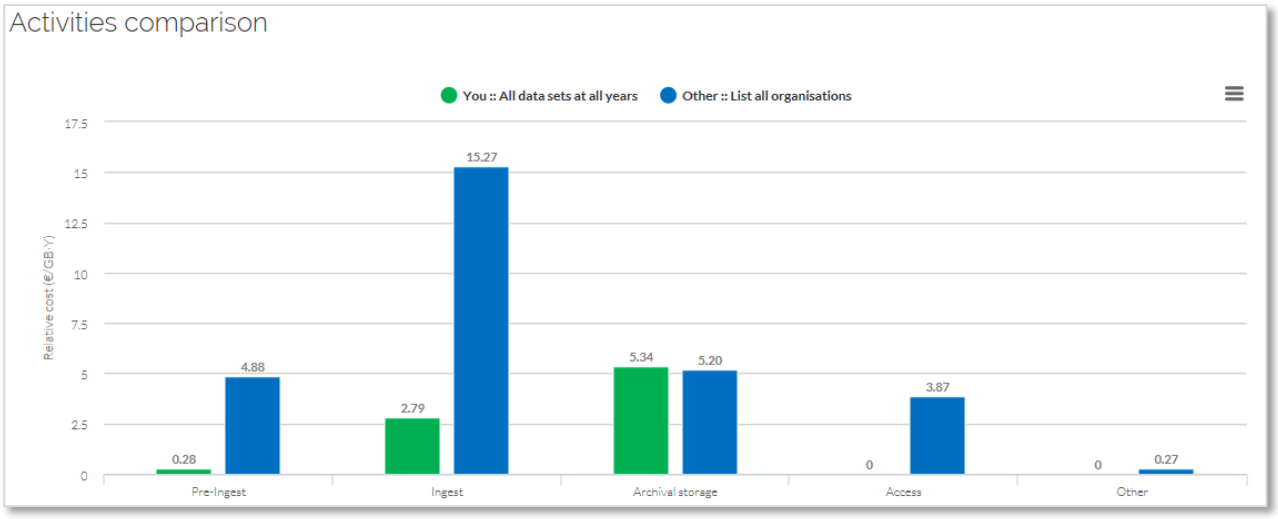


Figure 19—Graph displaying global comparison regarding activity-based costs

The 'Activities comparison graph takes an average total expenditure for all years and either compares an aggregated figure derived from all the user's costs (or the user's selected costs) with other cost data sets shared within the CCT.

All costs are converted to Euros (€) for the purposes of global and peer comparison.

It is possible to hover over each bar or use the key to identify the relative cost per gigabyte for the total period of the selected costs. The figure at the head of the bar for each year also shows the relative cost per gigabyte for the total period of the selected costs.

The graph below shows the same costs, but using costing by resource types.

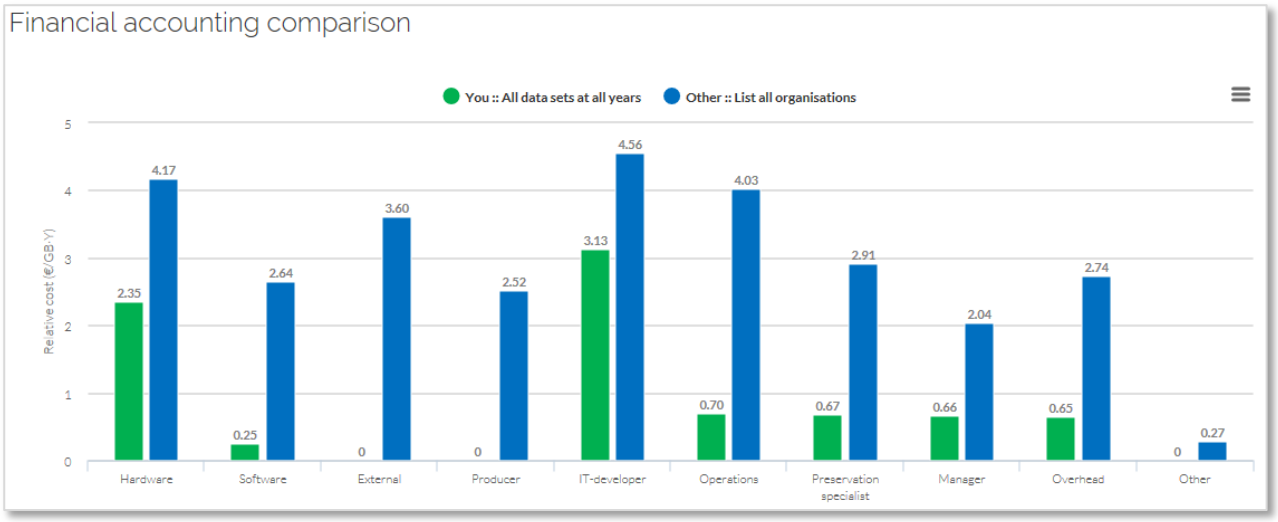


Figure 20—Graph displaying global comparison regarding costs based on costing by resource type

2.3.2.3 Peer comparison: Individual comparisons

The purpose of the peer-to-peer facility is to enable two similar organisations to compare their costs (for example, two memory institutions holding approximately the same data volume, employing the same preservation strategies, etc.). In some cases peers can contact each other to discuss process, address differences, share practices and knowledge—in brief, a communication that can lead to better use of resources.

The peer comparison offers:

- Manual selection of which peer to compare with
- An automatically generated 'similarities' filter allowing for quick identification of the resemblance of the chosen organisation
- The possibility of contacting a particularly interesting peer through the system

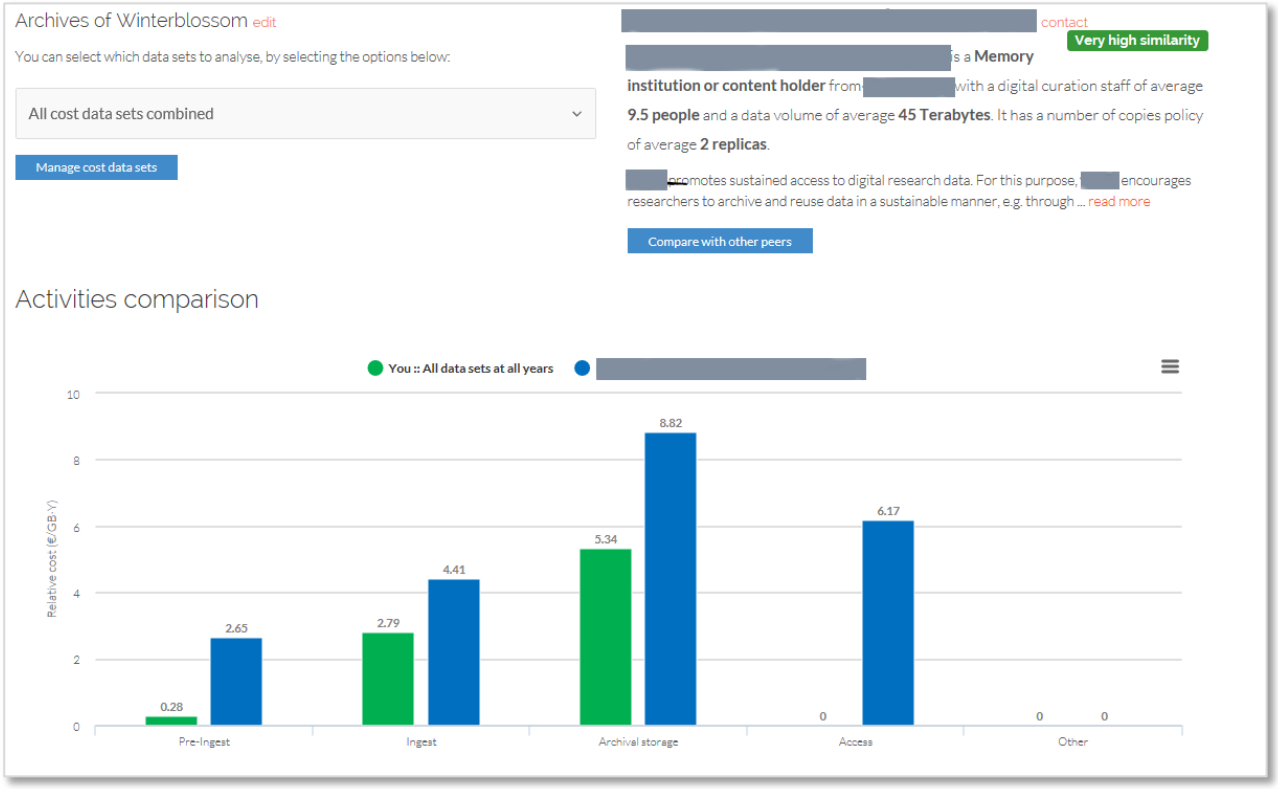


Figure 21—Graph displaying peer comparison regarding activity-based costs

3 Rationale and mechanisms

This section focuses on the most important mechanisms and principles that have guided the development of the CCT.

The two main—and somewhat contradictory—requirements for the success of the CCT are:

1. **Quantity of cost data**. The system need to have sufficient relatively recent data to make the aggregated results meaningful and statistically fit-for-purpose. To avoid discouraging users with overly complex and long winded data input processes the submission of cost data needs to be easy and quick.
2. **Quality of cost data**. In direct contrast to the need for simplicity outlined above, quality implies a more thorough data submission process. In particular, in order to make the output of the CCT usable, normalisation of the data is necessary. This could happen either during the submission process (by the submitting stakeholders themselves) or during the analysis process (by the administrators of the CCT) that follows. However, for reasons of accuracy[24], trust[25] and sustainability[26], it is considered to be important that the normalisation is done by the submitting stakeholders themselves.

Since the normalisation of cost data is difficult and time-consuming we have settled upon a compromise approach for the CCT—as described in section 2—where precision is partially sacrificed in exchange for a lower barrier to participation. Obviously there is a risk in this approach, particularly if the CCT cannot attract a sufficient number of stakeholders needed to achieve critical mass and sustain the tool beyond the duration of the 4C-project. This risk is being mitigated in part by the current effort to encourage stakeholders to use the tool as discussed in section 5—Outreach.

3.1 Overall rationale

Once a stakeholder has grasped the concept of the CCEx and in particular the CCT we feel that participation will be seen to be "a no brainer". Stakeholders will actively seek to be involved. The potential benefits vastly outweigh the effort required to participate at the most basic level and the additional benefits to be accrued from sharing data in a secure, trusted environment should also drive users to enter data and "go public". However, before this desire to be involved can be achieved, it is important that the initial reluctance to participate be overcome.

With this in mind we have put together a succinct list of potential benefits for using the CCT, the 'Overall rationale':

1. **Understand costs**—the CCT will help you better understand your costs through the process of submitting them
2. **Assess costs**—the CCT will help you assess your past costs based on activities and / or resource types (purchases and salaries)
3. **Compare costs**—the CCT will help you identify discrepancies in costs compared to global averages as well as costs of your peers

[24] The submitting stakeholder knows own costs better than we do.

[25] It is easier to convey a message of trust if the data manipulation is done by the owner of the cost data set than if a third party has handled it.

[26] It is important to lower the workload for the 4C-project so that the Cost Comparison Tool does not drain the 4C-project and so that the cost data submission can continue beyond the project lifetime.

4. **Clarify costs**—the CCT will help you communicate costs clearly (for example the costs of a project to a funder)
5. **Plan activities**—the CCT will help you understand and estimate the future costs (of , for example, a project an activity, etc.)
6. **Enhance communication**—the CCT will help you identify and communicate with relevant and interesting contacts
7. **Instigate collaboration**—the CCT will help develop a culture of sharing and collaboration within the community

3.2 Rationale and mechanisms of the cost data submission phase

The CCT enables stakeholders to upload cost data and the related cost information in order to create an output that can be used to fulfil one of the major needs articulated in the stakeholder consultation; that of obtaining efficiency by comparing costs, activities, and processes[27].

When one's objective is to enable the submission and normalisation of cost data it is crucial to acquire knowledge about the actual costing practices of organisations. This allows the submission process to be adapted to suit the user's data—to make the process as easy as possible for them—as opposed to forcing them to adapt to our process.

3.2.1 Research on existing costing methods

The two main costing methods in use today are costing by resource type[28] and activity-based costing (ABC)[29] respectively.

The costing method that the 4C-project has encountered the most during the investigations is *costing by resource type* (approximately 80% of those surveyed). For this reason, and to accommodate the majority of stakeholders, it is reasonable to offer the possibility of submitting cost data according to this method. Furthermore, this method does not exclude an activity-based breakdown of costs. If the stakeholders normalise their cost data onto an activity-based structure, it becomes possible to compare costs between stakeholders even though they use different costing methods. One problem is that stakeholders that do use costing by resource type quite often have considerable indirect costs that they qualify as 'Overhead', which makes the method less accurate.

Activity-based costing is the method that is the more accurate, because it is much more detailed. Furthermore, it excludes the use of 'Overhead' by assigning it to the direct costs and it specifies that we are in the field of digital curation, which makes it a pertinent method for this particular project. However, not many stakeholders surveyed use it, and when they do, it is often in a combination with the costing by resource type method.

The cost data that has been received from the 4C consortium partners also principally originate from the costing by resource type method in spite of the fact that many of the submitting organisations' core business is digital curation. This is because accounting is performed by the financial department, which is usually separate from the one engaging with digital curation, and that activity-based costing reports do

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

[28] This costing method breaks down costs in labour costs, capital costs and indirect costs.

[29] Activity-based costing breaks down costs in activities.

not conform to generally accepted accounting principles (GAAP)[30]. As such, organisations following activity-based costing methods would need to maintain two cost systems and accounting books, one for internal use and another for external reports, filings, and statutory compliance.

The 4C stakeholder consultation showed that in almost half of the organisations responding (48%) the costs of digital curation are not separated from other business activities, and 15% indicated that the costs are separated, but not further broken down. This complicates the submission process, because not only are the stakeholders asked to submit the cost data in a meaningful way by normalising them, but in order to do that, half of them need to separate the costs of curation from other activities in the organisation beforehand. Only 13% indicate that their costs are broken down into several activities, but not which, and the remaining 13% that they are broken down in several entities (e.g. departments, but it could also be into recurring or one-off costs, resource types, etc.).

Related to this, the stakeholder consultation also showed that it is rare for the person responsible for accounting and budgeting to possesses knowledge about the field of digital curation (at this question it was possible to tick several options, so proportions cannot be assigned[31]).

When the 4C consortium partners were asked to normalise their costs into a simplified activity-based structure the results corroborated the assumption that it is difficult and time-consuming to submit resource type costing onto an activity-based structure.

3.2.2 Normalisation

The only feasible way of making the cost data that exists in different organisations comparable is to normalise it. Normalisation in this context means adjusting values measured on different scales to a notionally common scale. It is used to bring existing cost data into conformity with a standard, in this case a standard invented by the 4C-project, but based on both de *jure* and *de facto* standards (see below).

3.2.2.1 Facilitating the normalisation process for the stakeholders

The assumption is that the best way of accommodating stakeholders is to adjust our systems to match their reality rather than the other way around. It was therefore decided to offer them the possibility of normalising their cost data according to their own practices: the ones using the costing by resource type method were allowed to normalise their figures onto a structure based on this method, and those employing activity-based costing were allowed to translate their figures to a normalised activity-based structure. This is why the cost data submission process integrates both options, as described in sections 2.2.1 and 2.2.2 above.

In order to lower the barrier for the submission of financial information, these normalised structures predefined by the 4C-project had to follow principles of *reduction of complexity* and *standardisation*. The reduction of complexity consisted in limiting the number of categories of the normalisation structure, and the standardisation consisted in drawing upon *de jure* and *de facto* standards when creating the normalised structures, both regarding terminology and concepts.

[30] http://www.brighthub.com/office/finance/articles/18891.aspx#ixzz0uFdq9vSV

[31] They are: General financial / accounts manager (34.78%); Department director (60.87); Repository manager (34.78); Asset owner (6.52%).

The normalisation categories for resource type costing

The normalised resource type costing categories that were chosen were therefore based on the resource type costing practices of the stakeholders (*de facto*) and incorporates the fundamental dichotomy of this methodology; the division between capital costs and labour costs.

In order to increase accuracy and detail, these two fundamentals were broken further down, and the field investigations lead us to use the sub-categories and terms outlined in Table 4:

Capital costs		Labour costs	
Sub-category	**Description**	**Sub-category**	**Description**
Hardware	Machines and media used throughout the whole digital asset lifecycle. Hardware may receive, store, validate, make copies, migrate and disseminate digital assets.	Producer	Any individual involved in creating digital content. This may include for example researchers generating and managing digital research data or aggregating new subsets of existing data for new analysis; government employees undertaking data collection and analysis. Producers may also include software developers and vendors who are producing code to enable analysis, manipulation and visualisation of digital content.
Software	Programs used throughout the whole digital asset lifecycle. Software may receive process, validate, create copies, migrate and disseminate digital assets.	IT developer	Staff members who develop software. Software engineers, programmers, system developers, coders.
External or third party services	Costs spent to buy services from 3rd party providers. Includes outsourcing, renting and leasing of hardware and software.	Operations	Staff members who execute technical tasks, for example testing digital material at ingest, operating the computers when migration occurs, burning optical disks, setting up robots, etc.
Overhead	All costs pertaining to overhead costs such as building costs, electricity, water, etc.	Preservation specialist	Staff members who execute the preservation planning of the managers; archivists who appraise digital assets, consult at access, execute administrative tasks.
		Manager	Staff members who organise and plan the work of digital curation in their organisation. Make tactical and strategic decisions, have staff responsibility and do budgeting.

Table 4—Normalisation categories for resource type costing

The normalisation categories for activity-based costing

The normalised activity-based costing categories that were chosen were based on the lifecycle of digital assets, which is *de jure* standardised in the OAIS reference model[32] and used in many other lifecycle models[33]. However, the community has to date never reached complete agreement on a common set of activities for curation, nor have they agreed on the naming of them. Keeping in mind the requirements of making the submission task easier for the stakeholders, it was decided to use the core lifecycle activities of digital assets, which implicitly means excluding what is considered as supporting activities[34]. The core lifecycle activities are identified as the reception, the storage and the dissemination of digital assets. All other activity costs are to be assigned to these core activities. After a lot of iterations testing these activity categories, it became obvious that it was necessary to add a fourth activity category pertaining to the preparation of digital assets for Ingest to meet the needs of important stakeholder groups, for example publicists, universities, big data and researchers. The normalisation categories for activity-based costing are shown in Figure 22:

Figure 22—Normalisation categories for activity-based costing

Table 5 below shows an example of how the normalisation process functions when normalising existing costs into the normalised activity-based costing categories. Cost units are displayed on the left-hand side and represent the detailed costs of a (fictitious) migration project. The activity cost categories on the right-hand side represent the predefined activity-based categories created by the 4C-project. Note that for simplicity this example only displays normalisations of the cost units Registering, Skills and Equipment (in bold).

Cost normalisation				
An organisations cost units and costs of migration project			**4C activity cost categories and aggregated costs**	
Cost unit name	**Cost (€)**	**Share %**	**Category**	**Cost (€)**
Pre-analyses	137,290		Pre-ingest	26,082
Scanning	69,631			
Registering	**13,548**	25%	Ingest	32,856
Main programmes	160,172	75%		
Migration	904,860	33%	Archival storage	46,274
Project Management	171,957	33%		
Skills	**68,086**	100%	Access	22,695
Equipment	**46,247**	33%		

Table 5—Example of cost normalisation

[32] ISO 14721:2012 Space data and information transfer systems—Open archival information system (OAIS)—Reference model http://www.iso.org/iso/catalogue_detail.htm?csnumber=57284

[33] For example the DCC Curation Lifecycle Model: http://www.dcc.ac.uk/resources/curation-lifecycle-model

[34] For example project management, metadata enrichment, scanning, preservation planning, etc.

In this example the imaginary stakeholder allocates the cost of “Registering” to the “Pre-Ingest” and the “Ingest” categories, but mostly to “Ingest” (75 %), because this is where most of the “Registering” costs were incurred. The stakeholder then splits the upgrading of “Skills” evenly between the activities “Pre-Ingest”, “Ingest” and “Access” (33 %). The stakeholder only purchased hardware (“Equipment”) for “Archival Storage” so 100% of the equipment costs are assigned to this activity category.

Again, the choice of normalisation categories in both costing methods follows the principles of reduction of complexity to meet user needs; and standardisation to meet the need for both comparisons and the obligation to be generic (to address all the identified stakeholder groups).

3.3 Rationale and mechanisms of the analyses and output phase

The analyses of the submitted cost data is automated and based on mathematical models and restrictions that are further explained in this section.

The analyses serve two purposes. The first is to be able to produce diverse results that are interesting to the stakeholders. The second—vital for the 4C-project—is to understand cost data, to learn how to structure it and to learn which parameters are needed in order to better understand and compare data as well as build cost concept model[35].

3.3.1 A word on precision and significant figures

There are a number of factors that affect the potential precision of the figures used in calculations. Due to the nature of the domain and the objectives of the Cost Comparison Tool, the accuracy of the submitted costs (which are mapped into normalised categories) as well as those generated as part of the comparison of costs with other groups and peers must be performed with limited precision and interpreted with caution. The information requested from the users should always be seen through the lens of ‘rough numbers’. The objective of the CCT is to guide users towards strategies that could potentially reduce the cost of their curation through giving a general idea of how they compare with others, and not to serve as a precise tool to evaluate and compare costs. It would perhaps be better to describe the figures as plausible as opposed to precise.

Examples of this reduction in precision are:

- the smallest available units for data volume and date intervals, which go down to Gigabyte and Year and not smaller.
- lack of support for partial numbers (only positive integers are supported) for data volume, date intervals, and costs
- the main mapping tool into categories is the slider, which goes into integer steps on a percentage scale

To avoid misleading precision, the processed information that goes back to the user is rounded to the maximum number of significant figures. As the least precise input data tool is the slider, which is used both on asset types distribution and on cost data mapping into categories, it was calculated that the maximum number of significant figures is 2. Therefore, all mapped costs and asset type data volumes are rounded to 2 figures (e.g. 12345 is rounded to 12000, 0.0012345 is rounded to 0.0012).

[35] http://www.4cproject.eu/d3-2-ccm-grs

3.3.2 Currency conversion and reasons for common currency on output

Costs can be entered using various currencies, as every organisation can select the currency they use on their profile. The costs are entered within a cost data set and cost data sets have defined date intervals. Conversion rates between two currencies change often but we have chosen to use a single rate for each year. In addition, conversion rates are not transitive. In other words, converting from currency A to B is not the same as converting A to C and then C to B. Due to these two issues, and to reduce error and provide coherent results to all users independently of their currency, the output results are always presented in Euros (which is the common currency for output results, selected by the 4C partners). Costs are converted into Euros when they are entered, providing to the users feedback as they are entering the information. The conversion into Euros is made using the average conversion rate for the years that the cost applies. Yearly conversion rates from supported currencies into Euros are defined on an administrative page within the system, and more currencies and conversion rates can be easily added. On the first released version, yearly conversion rates for British pound and US dollar for the last 15 years (from 1998) were added. If costs are added from a date interval for which no yearly rate exists, the rate for the closest year is used. If yearly rates are changed or added after cost data for that year is entered, converted values are re-calculated.

3.3.3 Concept and calculation of relative costs

To make the costs between organisations of different sizes comparable, the concept of relative cost must be introduced. This tries to provide a value that is independent of variables that significantly affect the value, such as data volume and the date interval applied to the cost data set. The currently selected relative cost definition is cost (in Euros) per Gigabyte per Year (€/GB/Y). T his value is calculated by dividing the cost value, converted into Euros, by the data volume in Gigabytes and then by the duration in years for which the cost data set applies.

3.3.4 Calculation of combined cost data sets

In the combined output results sections—My costs, Group comparison and Peer comparison—information from several cost data sets can be combined into a single distribution of relative costs of the defined categories. The costs to be combined can be filtered using the options of those areas. Mapped relative costs ($\bar{x}$) from different cost data sets (x) are combined using a weighted arithmetic mean formula (below), where the weight (w) is the data volume.

$$\overline{x} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n}$$

With this formula, costs from cost data sets with a bigger data volume have more importance on the final combined value.

On group comparison, costs from the selected groups are combined with the same formula using all cost data sets from all organisations that fit the selected group profile.

3.3.5 Calculation of peer similarity

The similarity score between peers, shown in the peer comparison area, is calculated by comparing a set of fields from organisation profile and cost data set information, each with a configurable weight that can be changed on the administration pages of the system.

The set of fields used and their default weight value are:

- Organisation type (weight 50)
- Data volume (weight 40)
- Main asset type (weight 20)
- Number of copies (weight 20)
- Staff size (weight 20)
- Scope (weight 10)

The weighted arithmetic mean formula is used to calculate and assign 'similarity' into the following levels:

- Super similarity (80-100% similarity)
- Very high similarity (70-80% similarity)
- High similarity (60-70% similarity)
- Medium similarity (40-60% similarity)
- Lower similarity (0-40% similarity)

For the organisation type, the weight of the field is multiplied by the percentage of match between types of the user organisation with the organisation that you are comparing. In this case the score is maximum, (or rather it is equal to the weight of the field) when the organisations have the same types.

The score of the data volume, number of copies and staff size fields is calculated in a similar fashion. The weight of the fields is multiplied by the percentage of difference[36] between the weighted average of corresponding field in both organisations. Similar weighted averages will result in closest scores to the weight of field.

The score of the main asset type and scope fields is calculated in a similar way. The weight of the fields is multiplied by the percentage of match between the set of values of field for each organisation.

The percentage of similarity between two organisations is calculated doing the ratio between the calculated score and the maximum score.

[36] The percentage of difference is the difference between two values divided by the average of the two values.

4 Confidentiality and anonymity concerning the submitted data

Anecdotally, one of the commonest reasons for not sharing data—even with a small group of "trusted" peers—is the fear that the data will become public and somehow harm the provider of that data; either reputational harm, or financial harm (or both) [37]. Feared negative impacts include cost comparison to the detriment of the provider in question ("provider x is significantly more expensive than provider y"), cost comparison without a full appreciation of the reasons for those costs (highly specialised, non-scalable curation costs more than large scale generic preservation for instance), non-comparable evaluations (provider x's cost per terabyte is compared unfavourable with provider y's cost even though they are not preserving the same type of content, are operating at a different scale and provide different levels of curation).

Despite the fact that some curation service provision has arisen from academia where there is a culture of sharing (of knowledge at least) the providers of the requested cost information within those organisations are generally financial officers. In that sub-community there most definitely is *not* a culture of sharing. Indeed, the exact opposite ethos holds sway. It is very rare indeed to find a financial officer who is willing to share (potentially) commercially sensitive information. Their default response is generally "No" without even considering any safeguards that might be in place.

Faced with such an apparently intractable problem the team had a choice of pursuing the information needed through Freedom of Information[38] type requests (some—but by no means all—within the target stakeholder group have a 'public body' remit) or through provision of safeguards that would reassure the entire community that a) their data would be safe, b) the system would allow them to limit the use and exposure of their data, and c) the system would ensure that comparisons are "fair".

The question of fair comparisons is considered elsewhere. This section concentrates upon the question of data security, confidentiality and anonymity.

4.1 Data security

There are a number of different classes of information handled by the CCEx and CCT including: personal information, business information, web content information and web management information. The information types are by no means unique and guidelines & methods for handling them securely in web environments are well established. The team use these methods[39] and follow the guidelines and are confident that the data is as safe as it can realistically expected to be.

4.2 Principles and terms

The key principles employed in the management of data provided by users of the CCEx are clarity, transparency, fairness and informed consent. We want people to use our tools. In particular we want

[37] The second most common reason for not sharing is the effort required to accrue the necessary information is seen as disproportionate to the potential benefit. In short, the "What's in it for me" message is not getting through. This is addressed elsewhere.

[38] Details relating to the UK Freedom of information Act and information requests can be found at https://www.gov.uk/make-a-freedom-of-information-request/the-freedom-of-information-act. Many other European countries have similar laws.

[39] Details of these methods will NOT be exposed here. One of the key guidelines is not to reveal too much about systems in place to safeguard servers and data.

people to be comfortable using our tools and to understand clearly the implications of clicking on any **Register**, **Save** or **Accept** buttons. These principles are outlined in the Terms & Conditions one the web site (which also incorporates aspects of a formal privacy policy). It is hoped that by doing so, one potential barrier to participation—that of confidence in the organisations integrity—will be minimised.

We want to be fair, transparent and have aimed to achieve ToS;DR Class A badges[40] across the board for all our terms and conditions.

As touched upon above, broadly speaking the terms and conditions and privacy policy relate to **personal information** (things like the users name, email address, and so on), **business information** (costs information for instance) and **use of the web site**. The key features of the policy are paraphrased in the sections that follow. The full terms & Conditions can be seen on the website[41]and the supporting Frequently Asked Questions (FAQ/Q&A) can be seen on the lower part of the same web page.

4.2.1 Personal information

This is the information that relates to a user personally. To make best use of the tools users need to register on the site. Many (but not all) of the tools and services require registration before users can access them. This is partially in order to allow us to communicate with users, but also to facilitate data update. In addition registration allows us to perform some basic identity verification.

Users are asked to provide basic information (name, email, etc.). Elsewhere they are also offered the opportunity to opt out of communications.

As far as personal information is concerned we emphasise within the policy that we will not pass on personal information to any third party without express permission.

4.2.2 Business information

The business information we gather relates to the type, location and mission of users organisations. We also collect and store information relating to the costs users associate with data curation (we call this the detailed source information). When users first set up an organisation to assign costs to we ask them to select the degree of openness they're comfortable with. The levels we have selected are:

- Allow the use of anonymised cost data to calculate averages[42] in the global comparison[43] result.
- Allow the use of anonymised cost data for peer comparisons.
- Allow registered users to make contact through the site.
- Allow cost data to be shared with registered users.
- Allow snapshots[44] of anonymised cost data to be collected periodically.

For business information we emphasise within the policy that we will not pass that detailed source information on to any third party without express permission.

[40] http://tosdr.org/ ToS;DR aims to creating a transparent and peer-reviewed process to rate and analyse Terms of Service and Privacy Policies in order to create a rating from Class A to Class E.

[41] http://www.curationexchange.org/index.php/60-privacy-policy

[42] Averages will always have at least 5 organisations.

[43] If this is the only box ticked data won't be used for peer to peer comparisons.

[44] In this context we mean capturing the data from the live system and storing it in an unchanging environment for later analysis.

4.2.3 Use of the web site

We're not selling a service and users are not buying a service so "service level agreements" are neither applicable nor offered. However, we recognise an implied commitment to keep the service alive and on-line and that commitment—on a best endeavours basis—is outlined in this section of the policy.

The CCEx is a community driven and enabled service. To be truly useful it depends upon users entering accurate information, preferably on a regular basis. This "obligation" is highlighted as part of the user's commitment to the whole process. We also point out the potential pitfalls of relying solely upon the outputs from the site to underpin business critical decisions.

The site has a forum for users' discussions. The project has neither the resource nor the will to moderate the content of those discussions, but at the same time we recognise the benefits of open discussion. Users are reminded to be respectful and consequences should they choose not to be.

The last part of the core section of the policy outlines our cookie policy (essentially we use cookie information for analytics).

4.2.4 Q&A

The Q&A section provides some answers to common questions users have about both the terms and conditions and the site as a whole. It is updated when required.

5 Outreach

As the success of the CCEx is so heavily dependent on user and community engagement, the 4C-project has developed a specific Communication Plan which aims to raise awareness of this resource and to encourage the exchange of data through the CCT.

This plan pertains particularly to optimising stakeholder engagement with the CCEx through the application of communications activities, and identifies the channels, audiences, information and content to be disseminated in order to achieve this objective.

It aligns key project messages with those specific to the CCEx, particularly highlighting the benefits of using the CCEx and anticipating any scepticism. The plan outlines the frequency with which communications will take place, milestones for communications, quality controls and performance indicators, as well as responsibilities for undertaking these activities.

5.1 Objectives and Purpose

The statement of principle behind the Communications Plan is to facilitate access to the CCEx, in the knowledge that information is shared without commercial exploitation, with a view to sustaining the currency and relevancy of this resource.

Derived directly from the 4C-project objectives, the plan and the activities identified therefore aim to meet this statement of principle. Specifically by:

- Identifying and encouraging organisations to share cost data in order to populate the CCT, thus sustaining the basis of its appeal to users (otherwise known as the 'Cost Quest').
- Identifying and engaging with organisations to take part in a CCT pilot with a view to shaping the resource to meet identified stakeholder needs.
- Communicating the benefits of using the CCT to 4C stakeholder groups in order to create demand for the resource, thus sustaining its usefulness and usability through ongoing use of the resource.

We are applying an 'open and social' communications model to engage users in sustainable dialogue throughout the lifetime of the project

5.1.1 Measures of Success

The success of the CCT related communications activities may be measured by the levels of:

- Initial stakeholder participation
- Sustained stakeholder participation

A good indicator of what we can expect is approximately the number of individuals who expressed their willingness to share cost information during the initial consultation (around 50). If just half of these individuals use the CCT, this could be deemed successful.

Awareness of the CCT and demand for the resource may be measured by the number of visits to the CCEx site.

Proposed success levels for the activities described are:

	Initial (Remainder of Project Lifetime)		Sustained (Post Project)	
	Cost data sets shared	Interaction with the CCEx (hits)45	Cost data sets shared	Interaction with the CCEx (hits)
Good	25	2000	5 per month	1000
Acceptable	12	1000	2 per month	500
Poor	5	500	Less than 1 per month	250

Table 6—Outreach success levels

During the 'Initial' period of measurement, it is anticipated that a quantity of 'test' and 'draft' data will be entered into the CCT. These measures refer to 'final' data sets only.

5.2 Key messages

Communications relating to optimising stakeholder engagement are being supported by the following proposition and key messages which highlight the benefits and value that may be derived by each stakeholder group. These messages should also be used to anticipate and combat any scepticism.

Drawing from use cases for of the each stakeholder groups, the benefits and key messages are listed in Table 7. Note that not all benefits concern just the CCT. Some relate to the whole CCEx as well. In other words, they relate to the whole website.

Stakeholder Group	Benefits
All	Any data shared through the CCT is done anonymously and confidentially, by aggregating inputs for organisation size and type and controlling any contact requests between organisations through the 4C-project team.
Research funder	The CCT enables the identification of activity costs which might be covered 'in-project' e.g. what/how much time is "reasonable," to ensure that funder guidance helps applicants to understand eligible costs, and that research funds yield the best return on investment.
Big data science	The CCT enables the identification of relevant costs in terms of domains, scale and activities, to assist with effective budgeting for new projects and making reasonable estimates from the outset.
Digital preservation vendor	The CCT shows real cost data related to operating a digital preservation service (either as a third party service or in-house solutions) and enablse comparisons of expenditure with that of competitors and peers, providing a greater understanding of the reasons for significant disparities in costs, and enabling the implementation of possible remediation.

[45] Website traffic measures were calculated using 4C partner, DPC's online statistics as an indicator

Stakeholder Group	Benefits
Government agency	The CCEx provides cost model descriptions that will help to inform which cost models to use. It enables comparison of expenditure between government organisations and assists with planning and budgeting.
Publisher or content producer	The CCT enables comparison of cost structures with those of other companies, highlighting areas of potential cost-reductions in day to day activities.
Data intensive industry	The CCT shows what large national organisations are spending to preserve access to collections so that access may be guaranteed to data for as long as it is required, whilst reducing risk in a cost-effective way.
Memory institution or content holder	The CCT offers controlled contact and comparisons between large and small institutions with a view to assisting in managing limited budget resources, creating business cases for future activity and benchmarking—with accuracy indicators to manage the different scales of operation.
Small or medium enterprise	The CCEx provides an assessment of different available cost models in order to identify which is the most suitable for a particular type of organisation.
University	The CCEx provides access to resources which will assist in meeting funders' expectations and raising awareness about Research Data Management and will help secure funding for additional resources.
Cost Model Experts	The CCEx enables users to find details about cost models and provide appropriate guidance so that refinements may be made based on additional feedback, making models more valuable for users
University	Using the CCEx Institutions are better able to plan for and sustain curation costs across the research lifecycle.
Researcher	Researchers can use the CCT to compare with other, similar projects; to see where costs could be brought down; to comply with funders' expectations, both regarding data management, preservation and expenses
Community member	The CCEx and CCT have prospects to grow and thrive as a community resource.

Table 7—Benefits for the stakeholder groups

5.3 Strategy and Timing

The CCEx communications plan employs a combination of push and pull communication strategies to encourage stakeholder participation with the CCT. These include:

- **Push strategy:** 4C is using its communications channels to 'push' the concept of the CCEx directly towards targeted stakeholder groups, supported by incentives and messages outlining long term benefits. This will include: personal invitation and direct contact through email/ phone; exhibitions and demonstrations; offer of incentives.
- **Pull strategy:** 4C communications channels are being used much more widely to convey the benefits of the CCT, creating a demand for participation and causing stakeholders to 'pull' more

information and opportunities to become involved. Pull strategies involve mass communications through: social media (Twitter, LinkedIn, etc.); PR via news releases and publications.

5.3.1 Timing

In order to optimise the participation in the CCEx development, it is anticipated that communications efforts should be focused within the following timeframes:

Months	Focus of effort	Strategy
12 to 21	Creation of demand	Pull
12 to 21	Encouraging sharing of data—population of the CCEx Shaping/testing prototype	Push
21 to beyond project completion	Sustaining demand	Pull

Table 8—Outreach time frame

Means and Method

All communications includes the key messages outlining the benefits of engaging and sharing data with the CCT as outlined in Section 5.2, Key messages.

5.3.2 Push strategy

Personal communication through:

Means	Method	Timing (Month)
1. Personal invitation and direct contact through email/phone	• Identify contacts within the 4C-project CRM who have expressed willingness to be involved with the CCEx and/or share cost data • Develop email templates for contact with existing and new contacts • Contact all individuals who had previously expressed willingness to share cost information in the initial consultation, and invite them to do so • Contact individuals who have previously expressed interest in testing and shaping the CCEx, and invite them to do so • Follow up all individuals who had previously expressed staying in contact with the 4C-project following the initial consultation and invite them to share data / test the CCEx • Request all 4C Partners examine their own contact lists, and those of any related organisations, with a view to identifying potential contacts for an invitation to share data / test the CCEx • All 4C partners to invite and follow up identified contacts to share data / test the CCEx	13—ongoing / maintain dialogue, once established

2. Exhibitions and demonstrations	• Approve timeframe for pilot. • Based on feedback from the Advisory Board, identify events which target research funders, research organisations and memory institutions within the relevant time periods • Contact attendees at the appropriate events to encourage participation in data sharing and CCEx testing. Potential events include[46]: ○ CENL, Moscow, Russia, May ○ IASSIST-Conference, Toronto, Canada, June 3-6 ○ ElPub, Thessaloniki, Greece, June 19-20 ○ LIBER Conference, Riga, Latvia, July 2-5 ○ IFLA, Lyon, France, August 16-22 ○ JCDL/TPDL, London, UK, September 8-12 ○ RDA, Amsterdam @ DANS, September 22-24 ○ iPRES, Melbourne, Australia, October 6-10 ○ ICA 2014, Archives & Cultural Industries Conference, Girona, Italy, October 11-15 ○ DLM (Data Lifecycle Management), Lisbon, Portugal, November ○ APA, Brussels, November ○ IMCW, Antalya, Turkey, November 24-26	13 to 22
3. Offer of Incentives	• See below	ongoing

Table 9—Outreach push strategies

5.3.2.1 Offer of Incentives

Despite understanding the benefits, some stakeholders may still be unwilling to share data (for example for reasons of complexity or confidentiality). The key messages outlined in section 5.2 above may be used in response to some of this scepticism, but further incentives may also be offered where appropriate. The following incentives/ counter arguments to encourage cost data exchange and engagement with the CCT are being considered:

- Assure all contributors that data may be shared anonymously
- Offer to visit stakeholder to work through data available if too complex to breakdown independently
- Offer early access to shared cost data
- Offer immediate access to sector specific information on cost modelling
- Opportunity to build networks and collaborations between similar organisations/institutions on basis of the CCEx

[46] At the time of publishing of this document some of these events are in the past. However, the original communications plan for the CCEx was drafted and put into practice at the beginning of the year (2014)

5.3.3 Pull

Mass communication through:

Means	Method
1. Social media	• Twitter • LinkedIn
2. PR	• Mailing lists • Website news releases: • Publications

Table 10—Outreach pull strategies

5.3.3.1 Social Media

Aligned with the 4C-project Communications Plan, and in order to reach a wide audience, the project is using the 4C website and blogs, plus Twitter feed and a #4ceu hashtag to publicise and create demand for the CCEx, whilst encouraging debate and participation.

Twitter

In order to raise awareness of the CCT and to encourage data sharing, maintain a continuous conversation about the tool, and create interest for more information, the following tactics are being used:

- **Updates**—content driven weekly updates, which link to blog posts on the 4C and other websites on the topic of the CCEx (outlining benefits and encouraging participation) and/or a CCEx section on the 4C website which maps progress, provides (controlled) access to the prototype for testing, screen shots/wire frames and commentary.
- **Discussions**—weekly/monthly twitter conversations either at planned and publicised times, or 'impromptu' conversations started by friends of the project asking pertinent questions about the benefits sharing data through the CCT.
- **Hashtag**—establishment of the hashtag #CCEx to enable those interested to follow updates and discussions.

LinkedIn

- **Discussions**—The CCEx may also be advertised through the LinkedIn Groups of related projects such as APARSEN. A specific 4C/CCEx LinkedIn[47] Group will not be established in order to maintain the ethos of an 'open and social' project.

5.3.3.2 PR

The following tools are being used to promote 4C and the CCEx activities.

- **Mailing lists**—target audiences are being reached through email lists for both the 4C-project and those of related EU/other projects. This channel is best suited for inviting attendance at events where the CCEx will be exhibited and demonstrated, or for inviting volunteers to share cost data and for prototype testing.

[47] LinkedIn groups are thought to be too closed to sit comfortably with the projects open aspect.

- **Website news releases**—News items are being published on the 4C and CCEx websites as well as released mailing list and social media. News releases are added to the news section and tweets are used to publicise new articles. This channel is be best used for raising awareness of activities and events where the CCEx will be exhibited and demonstrated. Websites to be used include:
    - 4C News pages
    - Project partner news pages
    - Related project news pages
- **Blogs and website**—aligned with the 4C-project Communications Plan, project partners are continuing to contribute blog posts in order to maintain an enduring web presence and awareness of the project. Blog posts on the progress of the CCEx development are being made on a monthly basis. Their purpose is to spark discussion and debate by asking or by tackling a controversial topic. The blog provides a commentary on the progress of the CCEx and will also feed into twitter.
- **Publications**—periodic partner and related project newsletters are also being used to report progress on development of the CCEx. Publications describe the benefits of participation and encourage cost data sharing. Opportunities in the national presses will also be sought.

6 The link between the Cost Comparison Tool and the Curation Costs Exchange

The CCT sits in the heart of the Curation Costs Exchange platform (http://www.curationexchange.org) and benefits from the synergy that is created by this cohabitation.

As we have seen, the CCT enables the sharing of sensitive cost data, providing the opportunity to identify greater efficiencies, better practices and valuable exchanges of information between peers. The tool also addresses an acknowledged reluctance to share this data by anonymising and amalgamating it for the benefit of comparison. Only if an organisation wishes to share its identity will the tool reveal these details.

Acknowledging that some organisations are not in a position to share digital curation costs—they may not even be at the stage of having any costs to share—the CCEx also draws together a wealth of other information designed to help users understand more about digital curation costs.

Whether organisations are planning to build a business case for investments in digital curation, carry out a cost appraisal or analysis, or want to know about cost modelling, a toolkit of resources drawn from the 4C-project is available and aims to provide that grounding through the lifecycle of digital curation costs[48].

[48] 'Understand your costs', http://www.curationexchange.org/make-the-case. Note that the URL ('make-the-case') does not yet reflect the new name of this section of the website, i.e. 'Understand your costs'.

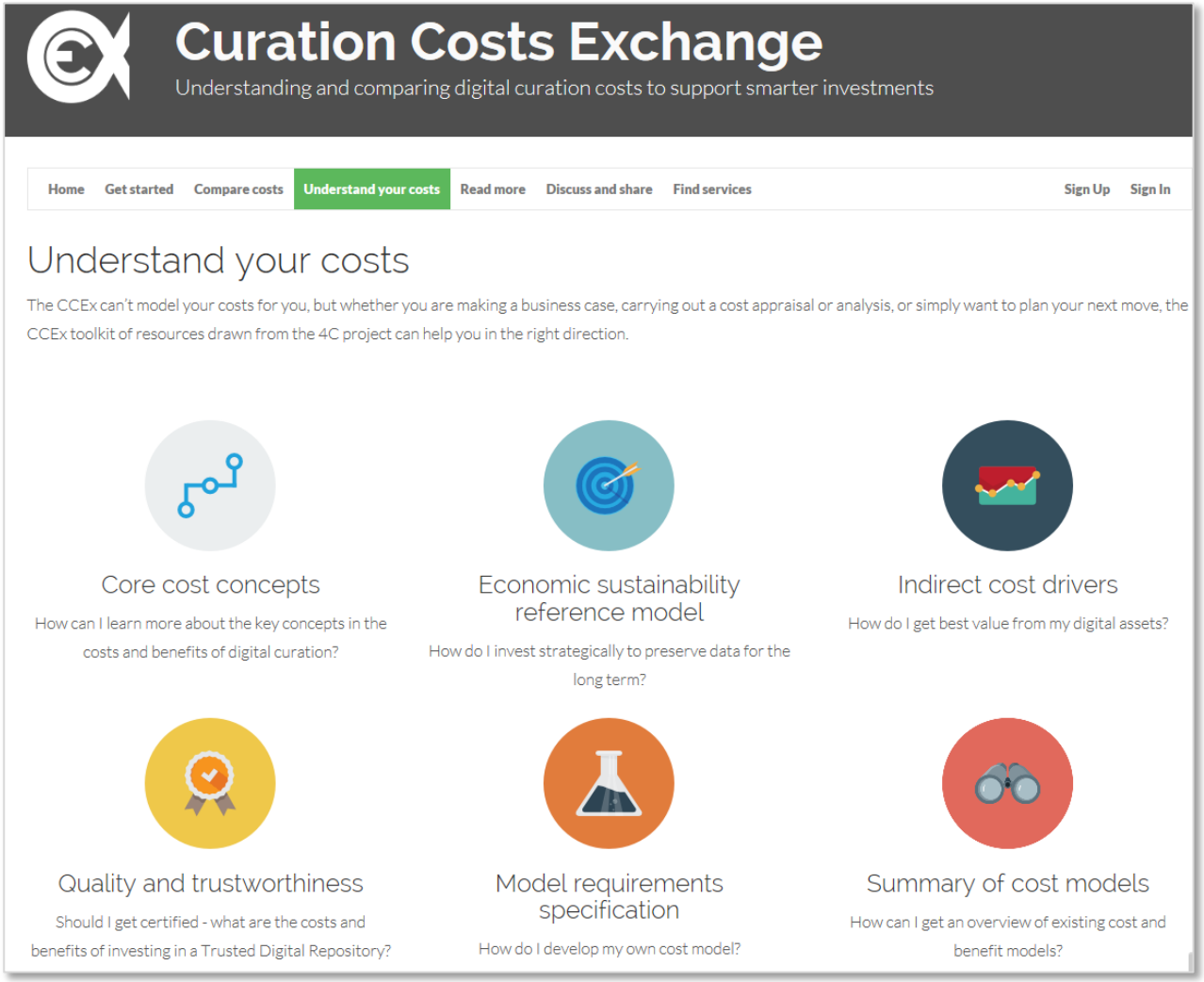


Figure 23—The CCEx 'Understand your costs' section

Core cost concepts[49] demonstrates the key concepts in the costs and benefits of digital curation; helpful when putting together a business case for spending of any kind in this area whether that is on staffing, activities or hardware and software. Linked to Core cost concepts, **Indirect cost drivers**[50] pin points a set of values significant in the practice of digital curation today. These values may incur cost but these costs will either realise a specific benefit or mitigate a particular risk. Applied at organisational management level, the drivers and related risks and benefits can help to inform decision making about curation investment and help to shape business cases, sustainability strategies and derive best value from digital assets.

The Economic sustainability reference model[51] highlights key digital curation concepts, relationships, and decision points in a complex problem space, helping users to benchmark and compare their own local models and invest strategically to preserve data for the long term. And for those considering certification, **Quality and trustworthiness**[52] examines the costs and benefits of investing in a Trusted Digital Repository.

[49] http://www.curationexchange.org/make-the-case/14-cost-concept-model

[50] http://www.curationexchange.org/make-the-case/16-indirect-cost-drivers

[51] http://www.curationexchange.org/make-the-case/15-economic-sustainability-reference-model

[52] http://www.curationexchange.org/make-the-case/17-quality-and-trustworthiness

Drawing from the 'Evaluation of Cost Models and Needs & Gaps Analysis' deliverable (D3.1) and providing an overview of selected cost and benefit models, the **Summary of cost models**[53] helps users identify which model may best suit their needs. Following the summary, if users still want to create their own cost model, the **Model requirements specification**[54] provides concepts and a generic specification for this task. The supporting information provided by the framework adds significant value to the comprehensive engagement by the 4C-project with various user communities and all of the detailed analysis of the requirements, drivers, obstacles and objectives related to that engagement.

In conclusion, this toolkit of resources puts the incurred digital curation costs of organisations into perspective, helping them to understand them in a more holistic way and plan accordingly.

The CCT captures past costs, and is not a means of cost modelling / projecting future costs. Since such functionality is of a highly complex nature[55], it has been deliberately left out of the CCT and incorporated in a more qualitative way in the different resources that can be found within the 'Understand your costs' toolkit.

Whilst the supporting information contained within the CCEx is of broad usefulness to the community it is at its most valuable when used by stakeholders to interpret results provided by their submissions to the CCT.

[53] http://www.curationexchange.org/make-the-case/19-summary-of-cost-models

[54] http://www.curationexchange.org/make-the-case/18-gateway-requirement-specification

[55] Note that all the cost models that have been evaluated by the 4C-project integrate this functionality, the success of which can be debated, cf. deliverable D3.1—Evaluation of Cost Models and Needs & Gaps Analysis, http://www.4cproject.eu/d3-1.

7 The future of the Curation Costs Exchange

The future of the CCEx as described in the 4C Draft Sustainability & Benefits Realisation Plan[56] (D1.1) focuses upon the need for 2-3 organisations being willing to take over the stewardship of the CCEx. Stewardship in this context implies website care, user care and data care and further development.

The negotiations with the organisations that have expressed a wish to take over the CCEx platform are on-going, and the elements of an agreement will only be finalised in time to be included in either the final CCEx deliverable[57] or the Final Sustainability & Benefits Realisation Plan (D1.2). Both are due in Month 24 of the project[58].

Possible future developments include:

- implementing cost predictive features (e.g. retention periods)
- creating more nuanced cost determinants (e.g. better differentiating between data complexities and data volumes) in order to allow for more detailed analyses of the submitted cost data
- establishing concrete integration between the CCT and the other resources of CCEx (see section 6)
- developing various dynamics to enhance user interaction

In order to secure the long term future of the tool, it has also been suggested that it should be incorporated in business analysis tools or be touted as a useful estimator for research grant application processes.

[56] http://www.4cproject.eu/community-resources/outputs-and-deliverables/d1-1-draft-sustainability-benefits-realisation-plan

[57] D2.8 Curation Costs Exchange.

[58] End of January 2015

8 Conclusion

The digital curation community recognises the potential of the CCEx, and in particular the CCT. The community welcomes a tool that will demystify the costs of digital curation and will bring about greater cost efficiencies.

The CCT is the mainstay of the CCEx platform. It provides the functionality that makes the website different from other websites promoting similar topics. The degree of success of the CCEx platform will thus not only be reflected by the general interest in this topic—which until now has been quite underexposed and underestimated—but it will also be measured in the readiness of stakeholders to share sensitive financial information. To a great extent this hangs upon the CCT's adequacy and pertinence: Does it work properly? Is it easy to use? Are the results interesting?

One way of achieving adequacy and pertinence has been to aim for a very high degree of user involvement in the development of the tool, and also to perform in-depth examination of user behaviour around the costing methods of digital curation. With this in mind, the Cost Comparison Tool is a product of the user requirements that the 4C-project has identified through its deliverables[59] and numerous user consultations and iterations in the form of workshops, webinars, focus groups, beta-testing, individual usability tests, Advisory Board consultations, internal meetings, poster sessions, and conference presentations.

Knowledge on how to create the best cost data submission flow has been amassed via desktop research, examinations of current accounting practices and standards and via inspiration from other projects[60], which have dealt with similar issues.

The development of the CCEx platform and of the CCT is ongoing, and, as such, this report represents but a snapshot of its functionality.

Future potential developments essentially revolve around extending the CCT's capacity to enable the estimation of future costs; enhancing the pertinence and impact of the cost determinants to allow for better analyses of the cost data; and augmenting user interaction.

Creating awareness of the costs of digital curation is also an important part of the potential impacts and benefits of the CCEx. It is broadly acknowledged that one of the challenges of the digital curation community is making funding bodies aware of the actual costs of digital curation and of the fact that these will not diminish as data volume explodes.

The two key challenges of the CCT remain the gathering of enough cost data and the translation of these into figures that organisations of all kinds find meaningful and engaging. The meaningfulness and usability of these figures is not only dependent on how they are normalised into comparable cost data (the quality of the data output), but also on how many cost data sets the CCT is able to attract in order to generate statistically reliable and useable figures (the quantity of the data output). At its inception the platform is faced with a potential of a catch 22 situation. The most compelling argument for the submission of cost data sets is a critical mass of useful, comparable data. But in order to achieve critical mass users must submit data to a system without critical mass, an altruistic gesture that some find hard to justify.

[59] Essentially deliverable D2.1 Baseline Study of Stakeholders & Stakeholder Initiatives and deliverable D3.1 Evaluation of Cost Models & Need & Gap Analysis.

[60] Especially the EU-funded APARSEN project which produced a deliverable on cost parameters for digital repositories (http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2013/03/APARSEN-REP-D32_1-01-1_0.pdf)

However, we feel that the CCEx and CCT together make up a formidable resource for the community and have instigated an active engagement programme to ensure that the community cannot fail to agree.

References

APARSEN (2013a). D32.1 Report on Cost Parameters for Digital Repositories

APARSEN (2013b). D32.2 Report on Testing of Cost Models and Further Analysis of Cost Parameters

DCC (2009). The DCC Curation Lifecycle Model
http://www.dcc.ac.uk/sites/default/files/documents/publications/DCCLifecycle.pdf

InterPARES (2007). Chain of Preservation (COP) Model http://www.interpares.org/ip2/ip2_models.cfm#

ISO 14641-1:2012 Electronic archiving — Part 1: Specifications concerning the design and the operation of an information system for electronic information preservation

ISO 14721:2012 Space data and information transfer systems -- Open archival information system (OAIS) -- Reference model

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. Vol. 12 (4), pp 195-214. http://link.springer.com/article/10.1007%2Fs00799-012-0092-1