Cost-Benefit Advocacy Toolkit

Costs Factsheet

Introduction to costs

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

A basic understanding of budgets and how different factors affect digital preservation and curation costs are critical to establishing and developing any data archive.

However, an understanding of the costs of preserving and curating research data sets is not enough in isolation for effective advocacy or to assess economic sustainability.

Cost analysis should be accompanied by an analysis of the anticipated benefits. This costs factsheet should therefore be read and used in conjunction with other components in the Cost-Benefit Advocacy Toolkit, particularly the Benefits Factsheet and the Return on Investment Factsheet.

Effort required and our knowledge-base

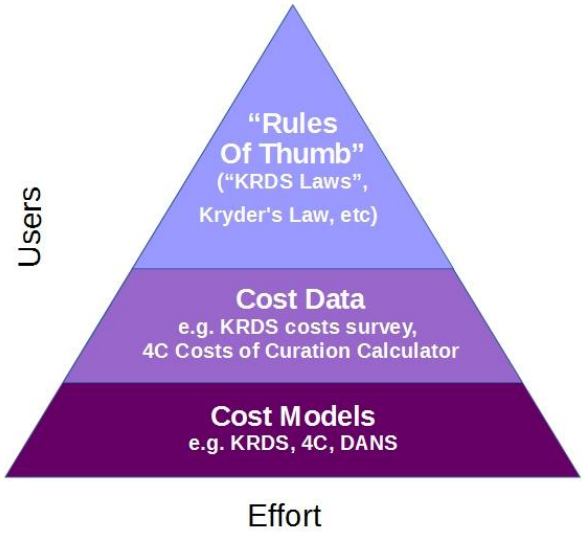
The costs of data curation and digital preservation have been the focus of a range of research projects in recent years and a selection of tools and a body of knowledge has emerged.

Costs are not a simple topic and in practice can be very complex. Costs in any organisation may be distributed across many departments, activities and budget headings. Establishing costs can therefore involve speaking to many different people and costs can be difficult to untangle. In addition, data curation costs are variable according to a range of economic and service factors that may be included/excluded. Issues such as inflation/deflation, cost of capital, depreciation, and scope and the levels of service provided, all affect costs.

This complexity means that the effort threshold for some costing activities such as detailed activity-based costing is very high and therefore direct use by individual data archives may be limited.

Project funded by the EU Horizon 2020 Research and Innovation Programme under the agreement No. 674939

The Toolkit uses a tripartite pyramid (Costs Models, Cost Data, "Rules of Thumb") as a means of understanding existing work, each building on (and requiring the existence of) the other in terms of a knowledge-base, and each requiring different levels of effort, and therefore ease of use. In terms of effort, costs models are the most demanding, cost data and then rules of thumb progressively less so. The reverse situation applies for use, with "Rules of Thumb" easiest to apply and potentially most widely used and cost models the least. Each of these groups is described in greater detail below.



Effort and Use Knowledge Pyramid: Costs Example. Charles Beagrie Ltd ©2017. CC-BY licensed.

Cost models

Individual cost models take many person-months of effort to research and build. Learning and adapting from previous work, or adopting it wholesale reduces the workout but it will still be significant. Only larger data archives or large (project) consortia involving smaller archives are likely to develop or apply them.

A 4C project research report (Bøgvad Kejser et al. 2014) provides an analysis of existing research related to the economics of digital curation. It outlines a basic terminology and general description of the components of cost models, and then provides a summary of ten cost models which it evaluates.

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

List of models identified as relevant to the field of digital curation (Bøgvad Kejser et al. 2014 Table 2)

The 4C analysis covers a wide range of sectors and tasks and only a small sub-set of these models will be relevant to social science archives: namely the Cost Model for Digital Archiving (CMDA) developed for DANS, the Dutch Data Archive; Keeping Research Data Safe (KRDS) a generic model, which has an implementation case study and input from the UK Data Archive; and potentially the 4C model itself, which is implemented in the Curation Costs Exchange.

Most of these cost models focus on costs of research data in an archive: few focus on pre-archive costs, the costs of data management activities before the point of depositing data with an archive. The UK Data Archive costing tool for data management may therefore also be of interest if you are developing costs guidance for research data creators.

DANS Cost Model for Digital Archiving (CMDA)

The CMDA model is a cost model for implementation in a specific archive – it was developed and customised specifically for use by DANS. It is an activity-based costing model utilising the OAIS Reference Model for its functional categories. It identifies activities and a set of costing components of each activity. It also takes the varying data complexity of datasets into account. Based on these factors the model estimates costs per dataset in "euros per dataset". Its results for DANS found that its datasets for archaeology are much costlier than its social sciences datasets, due to more variety and complexity in data formats (databases, images, geodata, CAD, etc). It also found that the CMDA activities "preservation" and "development of the archival system" were the most cost intensive, although this has to do with the expenditure on building up of the system which was still taking place at the time (4C 2014 p 30-32, Palaiologk et al. 2012, H.Tjalsma pers comm).

Keeping Research Data Safe (KRDS)

The KRDS activity model is a generic cost model for research data preservation with a user guide and template to support it being customised by its users for their specific institutional needs. Like CDMA it is a lifecycle activity costing model broadly based on the OAIS Model but with extensions to cover pre-Archive activities.

There is a "Lite" version and a "Detailed" version of the activity model. The "Lite" version provides a high-level granularity for allocating costs which could be sufficient to understand overall allocation of costs. This can be obtained with a much lower overhead in terms of capturing the required cost information (KRDS 2011).

The UK Data Archive was a partner in KRDS and provided a case study implementation as well as assisting with its overall development (Beagrie et al. 2009, 4C 2014 p 29-31).

UK Data Service Data Management Costing Tool and Checklist

The UK Data Service has developed an activity-based costing tool to help formulate research data management costs in advance of research starting, for example for inclusion in a data management plan or in preparation for a funding application (UKDS Data Management Costing Tool and Checklist 2014).

The tool considers the additional costs - above standard planned research procedures and practices - that are needed to preserve research data and make them shareable beyond the primary research team. The checklist indicates the activities to consider and cost to enable good data management. Such additional activities may require extra researcher or administrative staff time input, equipment, software, infrastructure or tools.

The 4C project

The 4C analysis of stakeholder requirements for financial information identified some important gaps and recommendations, including action to collect, describe, and exchange empirical cost data through a shared knowledgebase with cost data and use cases. 4C therefore developed a common framework for costs modelling that became the Curation Costs Exchange described below in Costs Data.

Costs data

Illustration by Jørgen Stamp
digitalbevaring.dk CC BY 2.5 Denmark

Costs data building on a pre-existing cost model can be easier to develop but still relatively complex if you need to compile it in new ways or levels of details for your organisation. You will need an understanding of cost models and underlying issues (see above) and OAIS functional categories (ISO 2012), to capture your cost data or to use/compare cost data from other organisations appropriately. The Curation Costs Exchange (CCEx) platform can assess the costs of curation practices through comparison and analysis. It provides guidance on preparing your cost data for this.

Curation Costs Exchange (CCEx)

CCEx was created by the 4C project and is hosted by the Digital Preservation Coalition as a community owned resource. CCEx provides comparison information of what you have spent (e.g. per gigabyte) on various activities, with similar institutions. You first submit profile information about your organisation, and cost data for specific activities (e.g. pre-ingest, ingest, storage, access) and projects. You add your data by year for each project/dataset/collection and say what it consists of (e.g. 10% databases, 25% video) and what you spent on different activities such as digitisation/preservation specialists etc. The tool then combines all your datasets for analysis. If you input your own information the tool will help find similar organisations so you can compare your costs either to the average costs of a group of organisations, or one-to-one with a single peer organisation.

All data that you submit is used solely for the purposes of building up aggregated data sets for comparison, and you can submit your data anonymously, even if you take up the option of contacting another organisation to discuss differences in spending and curation practices.

Keeping Research Data Safe 2 (KRDS2)

The KRDS 2 Identification of long-lived digital datasets for the purposes of cost analysis project, built on the work of the first "Keeping Research Data Safe" study and its activity cost model. It identified sources of long-lived data, developed longitudinal data on preservation costs, and analysed it to derive key factors affecting costs described below in Cost "Rules of Thumb". Thirteen cost datasets from 2009 and the survey criteria are available (KRDS 2009). The Curation Costs Exchange now provides more recent cost data sets but the historic cost data and approaches used in KRDS2 may still be of interest.

Cost "Rules of Thumb"

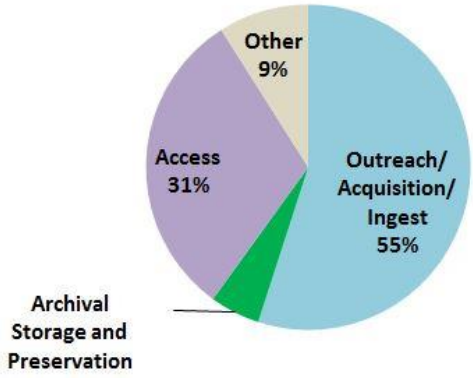
Cost "Rules of Thumb" or "Laws" are simple observations from existing cost data and projection of existing trends. These costs trends may hold for many years or even decades but eventually may alter: unlike laws of nature, which are fixed. They are very simple to apply and often very influential in business planning. "Moore's Law" and "Kryder's Law" have been critical in shaping development plans for industries in the IT sector. In digital preservation costs research generally, the major focus has been on developing cost models, and then gathering and comparing of cost data. However a general understanding of rules of thumb and trends within this work is likely to be useful to all social science archives, particularly those with fewer resources for gathering activity-based cost data or utilising cost models. Some of the key findings from the Keeping Research Data Safe (KRDS) research projects on digital preservation costs and details of Kryder's Law and Moore's Law are described below.

KRDS "Rules of Thumb"

Getting data in takes about half of the lifetime costs, preservation about a sixth, access about a third

KRDS found acquisition and ingest are the biggest costs over the preservation lifetime of research data. The costs of archival storage and preservation activities are consistently a very small proportion of the overall costs and significantly lower than the costs of acquisition/ingest or access activities for all the KRDS case studies.

Percentages varied between different archives but a consistent pattern emerged suggesting this rule of thumb from the Archaeology Data Service cost data as a rough guide to overall lifetime costs (Beagrie et al. 2010, pp. 31-52). It is potentially significant for those building business models and needing to fund archiving from depositor's research grants. Ingest costs may be within the timespan of the research grant and can be a significant part of lifetime costs.



Approximate Activity Data Costs for the Archaeology Data Service

(after Beagrie et al. 2010). CC-BY licensed

Preservation costs decline over time

KRDS found a trend of relatively high preservation costs in the early years reducing substantially over time for data collections. An example is the preservation costs projected for the Archaeology Data Service (ADS) based on their experience of the first 10 years of operating the data service. (Beagrie et al. 2008, pp.4-6).This long-term decline in costs reflects a number of factors: partly the effect of Kryder's Law on technical storage costs but mainly the growth in collections over time and the effect of economies of scale. Again it is potentially significant for those building business models, particularly if considering one-time fixed payment deposit fees or endowment for a dataset.

Fixed Costs are significant for most data archives

KRDS (Beagrie et al 2010, pp. 31-52) found that data archive costs are dominated by fixed costs that do not vary with the size of the collections. For most social science data archives, fixed costs such as core staffing and technical set-up will be significant.

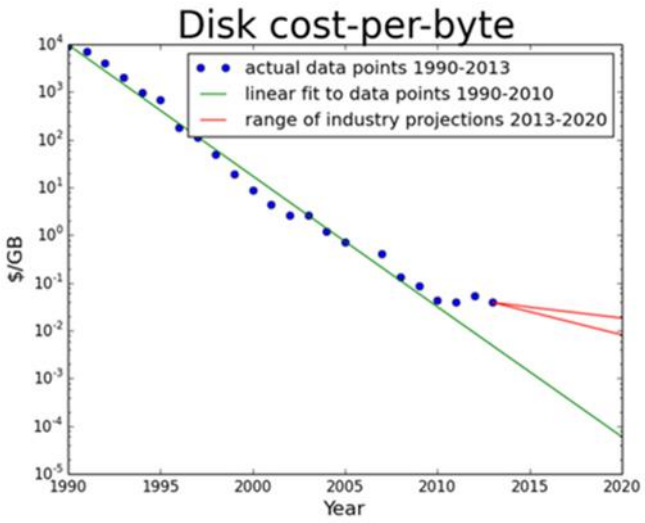
Fixed costs are eventually not fixed but you have to scale up quite a way before that applies. Activities characterised by significant fixed costs can reduce the per-unit cost of long-term preservation by leveraging economies of scale. These factors may have implications for cost-benefit of small collections (as relative costs can be higher) and for collection policies (economies of scale, lower costs and higher impact may come from collecting in adjacent areas such as population health data or the humanities, or via international data collaborations such as CESSDA).

Staff are the most significant proportion of archive costs

KRDS consistently found that staff are the major cost component overall, sometimes as high as 90% of the total costs (Beagrie et al 2010, pp. 31-52). This finding was also made in another recent costs study (NCDD 2017). Equipment costs are a relatively small proportion of total costs. There is a minimum base-level of staff and skills required for any service. It is important to note that staff are the most significant component of fixed costs (see above) and economies of scale will be largely driven by staff costs and data volumes.

Kryder's Law

Kryder's Law stems from an article published in Scientific American (Walter 2005) in which Mark Kryder observed that magnetic disk storage density doubles approximately every eighteen months. This has also meant that the cost of computer storage has roughly halved every eighteen months. It is a trend that has persisted over several decades but due to industry consolidation and greater costs of developing new technologies it may have begun to break down in 2010 (Rosenthal 2014, 2016).



Kryder slowdown. Chart by Preeti Gupta at UCSC (from Rosenthal 2014). CC -BY-SA licensed

In this graph the green line shows the projection of storage costs as they would be according to Kryder's Law. It shows that slowing started in 2010. The red lines are projections at the industry roadmap's 20% and a less optimistic 10%. If the industry projections continue, by 2020 disk costs per byte will be between 130 and 300 times higher than they would have been had Kryder's Law continued.

David Rosenthal points out the significance of this shift in Kryder's Law for preservation of data at scale, where storage has a major impact on costs (Rosenthal 2016).

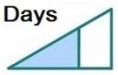
Moore's Law

Moore's Law refers to an observation made by Intel's co-founder Gordon Moore in 1965. He noticed that the number of transistors per square inch on integrated circuits had roughly doubled every year since their invention. From observation of this emerging trend, Moore extrapolated that computer processors would dramatically increase in power and decrease in relative cost year by year for the foreseeable future. Moore's prediction has proved broadly accurate for over five decades, and has been used to guide long-term planning and to set targets for research and development in the semiconductor industry. In 2015, The Economist newspaper suggested Moore's Law may be coming to an end. Transistors can still be shrunk further, but they are now getting more expensive. In addition cloud computing is making Moore's Law less relevant economically (The Economist 2015).

Linked toolkit resources	Effort
Benefits Factsheet, http://dx.doi.org/10.18448/16.0004	Hours
Return on Investment Factsheet, http://dx.doi.org/10.18448/16.0002	Hours
User Guide, http://dx.doi.org/10.18448/16.0001	Hours
Case study on using benefit and cost tools, http://dx.doi.org/10.18448/16.0006	Hours

Linked external tools

Curation Costs Exchange (CCEx), http://www.curationexchange.org/



UKDS 2014, Data Management Costing Tool and Checklist, http://www.data-archive.ac.uk/create-manage/planning-for-sharing/costing

KRDS Cost Activity Model (Lite and Detailed versions) and KRDS User Guide available from http://www.beagrie.com/krds/

Other references

Bøgvad Kejser U., Hougaard Edsen Johansen K., Thirifays A., Bo Nielsen A., Wang D., Strodl S., Miksa T., Davidson J., McCann P., Krupp J., and Tjalsma H., 2014, *D3.1 Evaluation of Cost Models and Needs & Gaps Analysis Final Report (Revision 1)*, http://www.4cproject.eu/d3-1/

Beagrie, N., Chruszcz, J., and Lavoie, B., 2008, *Keeping Research Data Safe. A Cost Model and Guidance for UK Universities,* http://www.webarchive.org.uk/wayback/archive/20140615221657/http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf

Beagrie, N., Lavoie, B., and Woollard, M., 2010, *Keeping Research Data Safe 2*, Final Report April 2010, http://www.webarchive.org.uk/wayback/archive/20140615221405/http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf

The Economist, 2015, *The End Of Moore's Law*. 4 Jan. 2015. http://www.economist.com/blogs/economist-explains/2015/04/economist-explains-17

ISO, 2012. *ISO 14721:2012 Space Data and Information Transfer Systems - Open Archival Information System (OAIS) - Reference Model*. Geneva: International Organization for Standardization. http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=57284

KRDS, 2009, *Keeping Research Data Safe 2 Project Website* http://www.beagrie.com/jisc/

KRDS, 2011, User Guide for Keeping Research Data Safe. Assessing Costs/Benefits of Research Data Management, Preservation and Re-use, http://www.beagrie.com/static/resource/KeepingResearchDataSafe_UserGuide_v2.pdf

NCDD, 2017, *Onderzoek naar de kosten digitale duurzaamheid*. Nationale Coalitie Digitale Duurzaamheid (Dutch Coalition for Digital Preservation), project report, http://www.ncdd.nl/kennis-en-advies/ncdd-publicaties

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*. Int J Digit Libr (2012) 12: 195. -1, doi:10.1007/s00799-012-0092 http://link.springer.com/article/10.1007/s00799-012-0092-1

Rosenthal, D. 2014, *The Half-Empty Archive*, DSHR's Blog, http://blog.dshr.org/2014/03/the-half-empty-archive.html

Rosenthal, D, 2016, *The Medium-Term Prospects for Long-Term Storage Systems*, DSHR's Blog, http://blog.dshr.org/2016/12/the-medium-term-prospects-for-long-term.html

Walter, C., 2005, *Kryder's Law*, Scientific American, https://www.scientificamerican.com/article/kryders-law