



**RDM: The costs of long-term
retention and access to research
data**

Matthew Addis, Arkivum Ltd

Webinar 3

© Arkivum Ltd 2013 **ARKIVUM**
RESEARCH DATA MANAGEMENT

Welcome to today's webinar on the costs of long-term retention and access to research data.

I'm Matthew Addis, CTO of Arkivum, and before that I spent 16 years at the University of Southampton on a series of research and development projects including digital preservation and cost modelling.

Webinar series	
Webinar1	Effective Long-term retention and access of Research Data
Webinar2	Long-term data management: in-house or outsource?
Webinar3	The costs of long-term data research data retention and access
<small>© Arkivum Ltd 2013</small>	
<small>ARKIVUM</small>	

This webinar is the third in a series of three webinars on Research Data Management. Previous webinars are available online if you missed them and we'll be making this webinar available too, including all the slides and a full set of speakers notes.

- The first webinar was an overview of the RDM landscape and issues.
2. In the second webinar, I was joined by Neil Beagrie and between us we explored of the pros and cons of in-house or outsourcing.
 3. In the third webinar today, I'll explore the costs of data retention and access.

If you've been on the first couple of webinars, you'll know I go through material pretty fast, and this webinar will be no exception!

Outline

Message: long-term RDM is not cheap!

- Costs v.s. benefits
- How to estimate costs
- Example costs and prices
- What Arkivum can offer

© Arkivum Ltd 2013

ARKIVUM
RESEARCH DATA MANAGEMENT

OK, so the message that I want to get across straight away is that long-term RDM can be expensive.

RDM has to be done properly, it's an active and ongoing process, and if you try to cut too many corners then you are at risk. There is the risk of research data not being captured properly in the first place, the risk of research data no longer being there when you need it, and at risk of research data not being in a form that you can find or use in the future.

So first of all sorry for the bad news!

But what we'll do in this webinar is look at how to estimate the costs of doing things properly, we'll try to put some real numbers against the costs, and then spend 5 minutes at the end looking at how Arkivum's solution is designed to help.

Costs v.s.benefits

- New research income
- Higher funding success rates
- Data centric publications
- Citations and research impact
- Reputation and kudos
- Cost savings
- Better reuse of infrastructure

© Arkivum Ltd 2013 ARKIVUM
RESEARCH DATA

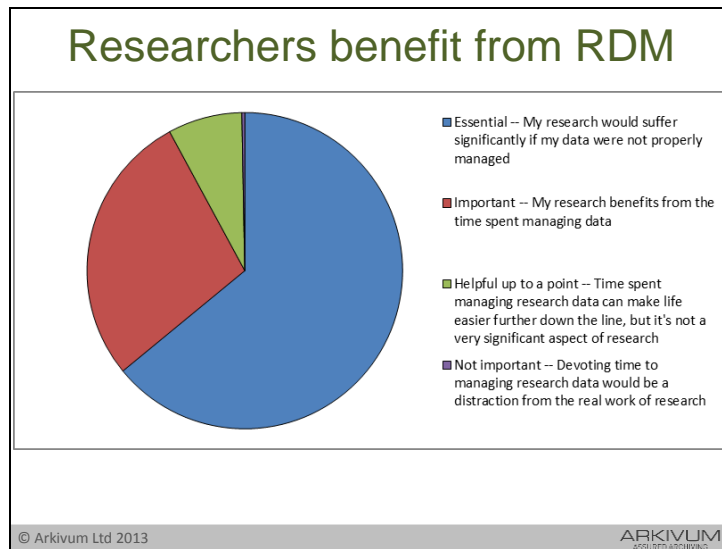
Talking about the costs of research data management, including long-term retention and access, doesn't make sense without considering them in context of the value and benefits of keeping data and making it available. The costs need to be proportional to what you want to achieve.

The benefits of retaining and making research data accessible can be both positive, for example helping to create a reputation for an institution or getting more research funding, but it can also include the benefits of preventing future costs e.g. recreating lost data, not servicing access requests, not meeting funding body requirements.

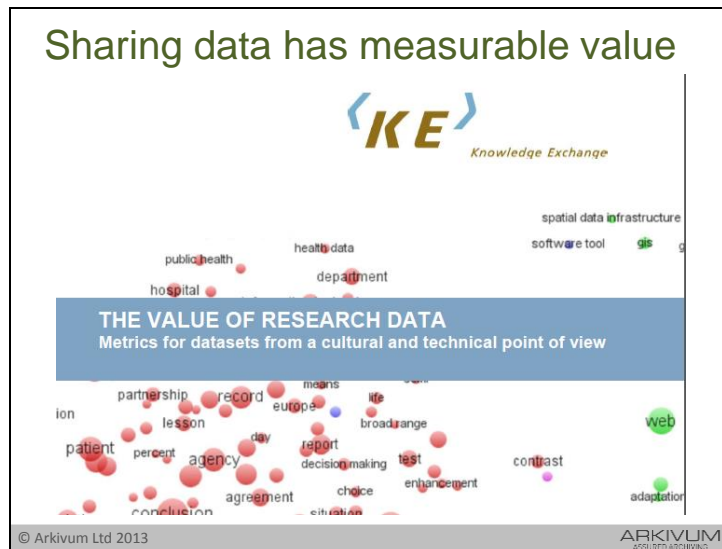
This webinar will focus a lot on cost, but you can go back to previous webinars where we talked more about value.

There's now a lot of evidence for the value of keeping research data safe and accessible. For example, the Keeping Research Data Safe project which provides lots of guidance on how to capture and describe the benefits of retaining research data.

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



Looking at some more specific examples - this pie chart shows some of the results from a survey done by Oxford as part of their Data Management Roll Out (DaMaRo) project. Researchers clearly see a benefit in RDM. In fact the vast majority see it as either essential or important.



And there is value in making that data accessible beyond the original researchers who created it.

The Knowledge Exchange project has looked at the measurable value of sharing research data,. This is something that I'll come back to later as one of the drivers from the funding bodies when we look at the challenge of cost recovery.

<http://www.knowledge-exchange.info/datametrics>

RDM: better, faster, cheaper

Measuring the Benefits

- 37%** Projected saving in staff time from moving Oxford University Classics Dept database to centralised virtual service (38)
- 69%** Increase in citations for clinical trial publications associated with making their microarray datasets publicly available (14)
- 500%** Growth in datasets downloaded from Economic and Social Data Service 2003-2008 (36)
- One-day delay cut to 5 minutes** Estimated time saving for crystallography researchers to access results from Diamond synchrotron, by deploying digital processing pipeline & metadata capture system (38)

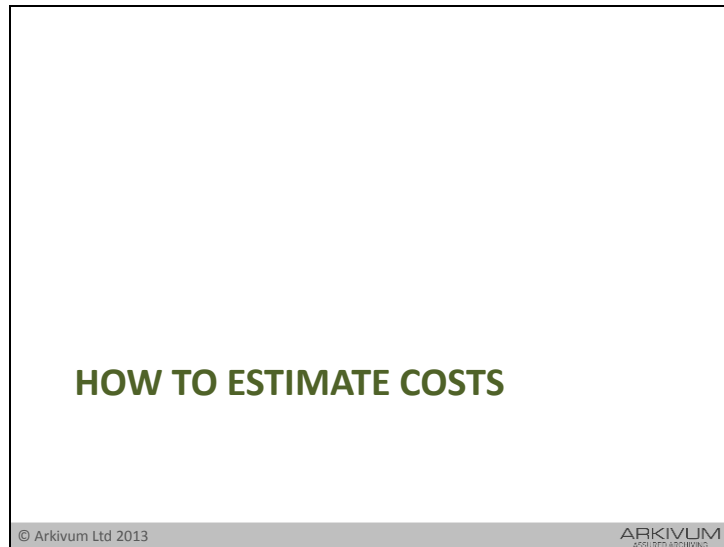
(See sources of further information)

© Arkivum Ltd 2013 **ARKIVUM** RESEARCH DATA MANAGEMENT

In fact there is a growing body of case studies that quantify benefits both short term and long term for RDM.

And benefits are the key to a business case for investment into research data retention. This is about having the inputs for a cost-benefit analysis – especially when the long-term sustainability of RDM needs to be assessed.



Slide 8



So let's get back to costs.

3 approaches

1. Historical data
2. Empirical models
3. Simulation



© Arkivum Ltd 2013

ARKIVUM
POLICY CENTRE

There are three ways to estimate costs.

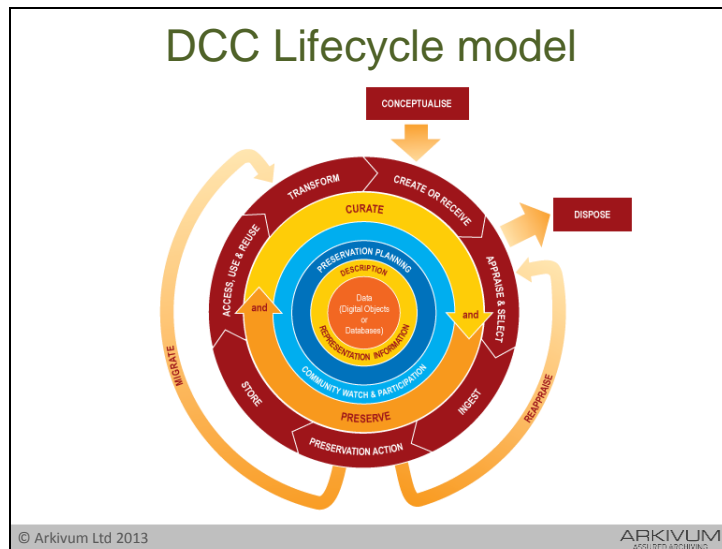
You can look at costs incurred in the past, and use these to forecast future costs

You can build models of all the areas that generate costs and use this as a framework for calculating the total cost, which is what empirical models are all about.

Or you can try to simulate how the world behaves and use that to predict what will happen given some starting point.

Basically, it's just like weather forecasting with a bit of crystal ball gazing thrown in.

Costs today are relatively easy to predict, costs tomorrow aren't too hard, but by the time you get to any form of long-term forecast it's often fingers in the air and can be anyone's guess.



Most people who've built cost models do so by looking at the lifecycle of digital content and all the activities involved as a way to break down cost analysis.

This slide shows the well known lifecycle model from the Digital Curation Centre, which provides a good start.

Using a lifecycle model helps ensure that nothing gets left out, but it also allows the boundaries of the cost model to be well defined.

<http://www.dcc.ac.uk/resources/curation-lifecycle-model>

Empirical models (LIFE)

<http://www.life.ac.uk/2/documentation.shtml>

$$\boxed{L_T} = \boxed{Aq} + \boxed{I_T} + \boxed{M_T} + \boxed{Ac_T} + \boxed{S_T} + \boxed{P_T}$$

L_T	Total cost
Aq	Acquisition cost
I	Ingest cost
M	Metadata cost
Ac	Access cost
S	Storage cost
P	Preservation cost.

The subscript T means that costs have to be calculated over the *lifetime* of the items being preserved.

© Arkivum Ltd 2013ARKIVUM
DIGITAL ARCHIVE

The LIFE model from the British Library is a good example of an empirical model based on a lifecycle. They defined each of the stages of the lifecycle: acquisition, ingest, metadata, access, storage and preservation and then worked out the cost of each one over time.

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

Slide 12

- Detailed breakdown into functional areas
- Examples
- Spreadsheets
- Guideline

Acquisition	Ingest	Metadata	Access	Storage	Preservation
Selection (Aq1)	Quality Assurance (I1)	Characterisation (M1)	Reference Linking (Ac1)	Bit-stream Storage Costs (S1)	Technology Watch (P1)
IPR (Aq2)	Deposit (I2)	Descriptive (M2)	User Support (Ac2)		Preservation Tool Cost (P2)
Licensing (Aq3)	Holdings Update (I3)	Administrative (M3)	Access Mechanism (Ac3)		Preservation Metadata (P3)
Ordering & Invoicing (Aq4)					Preservation Action (P4)
Obtaining (Aq5)					Quality Assurance (P5)
Check-in (Aq6)					

© Arkivum Ltd 2013

ARKIVUM
Digital Archive

There's a detailed breakdown of the activities involved and several case studies available on line of people who have followed the model, including real cost numbers.

<http://www.life.ac.uk/2/documentation.shtml>

Empirical models (CMDP)					
INGEST (IN)	ARCHIVAL STORAGE (AS)	DATA MANAGEMENT (DM)	ADMINISTRATION (ADM)	PRESERVATION PLANNING (PP)	ACCESS (AC)
Generate SIP	Receive Data	Administer Database	Negotiate Submission Agreement	Monitor Designated Community	Coordinate Access Activities
Receive Submission	Manage Storage Hierarchy	Perform Queries	Manage System Configuration	Monitor Technology	Generate DIP
Quality Assurance	Replace Media	Generate Report	Archival Information Update	Develop Preservation Strategies and Standards	Deliver Response
Generate AIP	Error checking	Archive Database Updates	Physical Access Control	Develop Packaging Designs and Migration Plans	
Generate Descriptive Information	Disaster Recovery		Establish Standards and Policies		
Co-ordinate Updates	Provide data		Audit Submission		
			Activate Requests		
			Customer service		

<http://www.ijdc.net/index.php/ijdc/article/viewFile/177/246>

© Arkivum Ltd 2013

ARKIVUM
POLICY CENTRAL

The Danish National Archive took a similar approach with their Cost Model for Digital Preservation, and again looked at the different functional areas of digital preservation and the activities in each one.

www.ijdc.net/index.php/ijdc/article/download/177/246

Format Interpretation (pw) =
number of pages * time per page (min) *
complexity (L, M, H) (* quality (L, M, H))

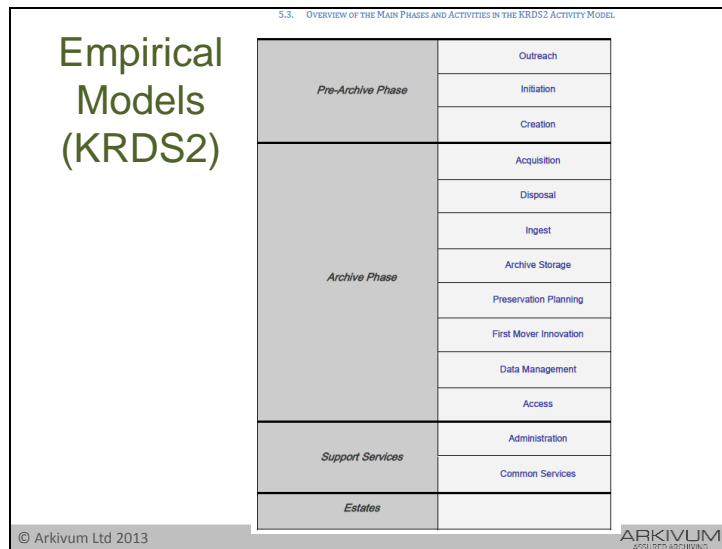
	Case 1		CMDP		CMDP - Case 1	
	pw	%	pw	%	Δ pw	%
IP Designs	44	12	50	24	6	12
A (1968-1998)	29	66	20	40	-9	-31
B (1999-2000)	15	34	16	32	1	6
C (2001-2004)	0	0	14	28	14	n.a.
B & C	15	34	30	60	15	50
Migration Plans	150	42	39	19	-111	-74
A (1968-1998)	105	70	15	38	-90	-86
B (1999-2000)	30	20	14	36	-16	-53
C (2001-2004)	15	10	10	26	-5	-33
B & C	45	30	24	62	-21	-47
Prototypes (Software Provision)	164	46	116	57	-48	-29
A (1968-1998)	101	62	48	41	-53	-52
B (1999-2000)	50	30	36	31	-14	-28
C (2001-2004)	12	7	32	28	20	62.5
B & C	62	38	68	59	6	9
Migration Package (total)	358	100	205	100	-153	-43

© Arkivum Ltd 2013 ARKIVUM
POLICY GROUP

But what's interesting for me about the CMDP work is that they looked at the cost involved in handling different file formats in the archive. Curation and preservation costs can vary enormously with the type of data that's being handled.

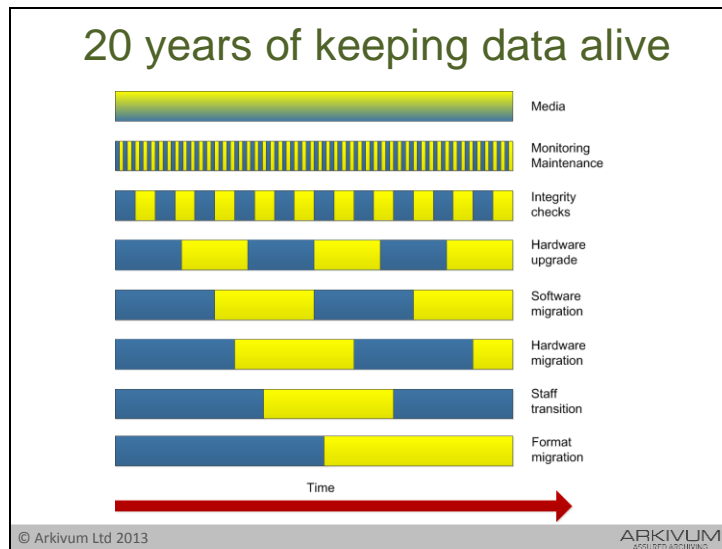
The CMDP approach is to model the effort required from staff in order to support different file formats. The model asserted that the effort was proportional to the complexity of the format as measured by the format specification document. The bigger and more complex the spec, the more effort it took to support and migrate the format in the archive. They then verified this assumption by using historical data. In other words this helped validate and calibrate the model, which is really important.

This also raises the important point that many of the costs of digital preservation are people related. This means staff costs should be one of the first ports of call when modelling costs.



Perhaps more familiar in the research data management space is the Keeping Research Data Safe project, which takes a similar approach to LIFE, CMDP and other projects in building an empirical model and then asking institutions what their costs were in each area. I'll be looking at the results of KRDS more later.

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



But what I want to do next is to look at how people model the long-term aspect of costs. This slide is one I used in previous webinars. This picture shows just some of the things that will happen over 20 years of trying to retain data. In the diagram, a change from blue to yellow is when something happens that has to be managed. In a growing archive, adding or replacing media, e.g. tapes or discs, can be a daily process, so is effectively continual. The archive system needs regular monitoring and maintenance, which might mean monthly checks and updates. Data integrity needs to be actively verified, for example annual retrievals and integrity tests. Then comes obsolescence of hardware and software, meaning refreshes or upgrades that will typically be 3 – 5 years, for example servers, operating systems, application software. In addition to technical change in the archive system is managing staff transitions of those who run the system, for example support staff and administrators. Even the format of the data being held may need to change because applications that read the data become obsolete or get upgraded and are no longer backwards compatible.

The picture is one of non-stop change – there's something happening all the time and it never stops. So, for the issue of cost, it's a question of properly accounting for the ongoing costs not just the start-up costs. And it's a case of making sure all the costs are considered, even the ones that might not be incurred until several years into the future.

Princeton model (1)

Pay Once Store Forever (POSF)

C = initial cost of the physical storage
D = rate at which the cost of storage decreases each year
r = number of years before storage must be replaced.
T = the total cost of the storage

$$T = C \times \frac{1}{1 - (1 - D)^r}$$

© Arkivum Ltd 2013ARKIVUM

Princeton estimated the long-term costs based on a model of regular migrations in work they did in 2009.

This established the concept of POSF.

Their formula is very simple. If you know the initial costs and the rate at which they fall, then you can calculate the total lifetime costs.


So suppose the cost halves every year. If the cost is £100 this year, then it will be £50 next, £25 the year after, £12.50 the year after than and so on. Add this up forever and you come to a single value, which is £200. So, if the costs halve every year, then the forever cost is simply twice the year1 cost.

So the approach becomes one of calculating the initial cost and then applying a multiplier.

http://dspace.princeton.edu/jspui/bitstream/88435/dsp01w6634361k/1/DataSpaceFundingModel_20100827.pdf

Princeton model (2)

- Applies to people, buildings, software
- Everything scales with data volume
- 20% decrease per year, 4 year cycle
- \$6000 per TB POSF (2009)



DATA SPACE

© Arkivum Ltd 2013

ARKIVUM
POLICY CENTRE

Princeton argue that this model applies not just to storage hardware, but also to people, space, power, software etc. because they are all in someway proportional to the amount of data being stored and hence follow the same trend downwards.


But their own use of their model is limited just to storage costs. They take a conservative view that the decrease in costs is 20% per year and that migrations for storage are needed every 4 years, which gave them a POSF of \$6000 per TB in 2009 (which by their own rules should now be down to \$3000).

<http://dataspace.princeton.edu/jspui/about/home.jsp>

California Digital Library

$$TCP = A + n \cdot P + m \cdot W + \ell \cdot C + k \cdot S + j \cdot M + i \cdot V + O$$

<i>TCP</i>	Total cost of preservation for all <i>Producers</i> .
<i>A</i>	Fixed cost of the baseline archival <i>System</i> .
<i>n</i>	Number of content <i>Producers</i> .
<i>P</i>	Unit cost of supporting a <i>Producer</i> .
<i>m</i>	Number of submission <i>Workflows</i> .
<i>W</i>	Unit cost of supporting a <i>Workflow</i> .
<i>ℓ</i>	Number of <i>Content Types</i> .
<i>C</i>	Unit cost of supporting a <i>Content Type</i> .
<i>k</i>	Number of units of preservation <i>Storage</i> .
<i>S</i>	Unit cost of <i>Storage</i> .
<i>j</i>	Number of preservation <i>Monitoring</i> activities.
<i>M</i>	Unit cost of a <i>Monitoring</i> activity.
<i>i</i>	Number of preservation <i>Interventions</i> .
<i>V</i>	Unit cost of an <i>Intervention</i> .
<i>O</i>	Fixed cost of administrative and managerial <i>Oversight</i> .

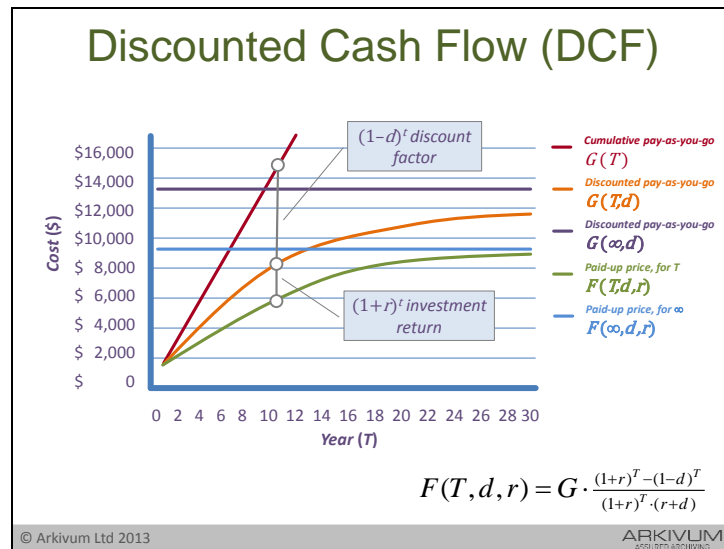
© Arkivum Ltd 2013 <http://wiki.ucop.edu/display/Curation/Cost+Modeling> 

The Princeton model has been followed by several UK Universities in estimating their internal long-term costs, especially for storage.

But it's a very simplistic model.

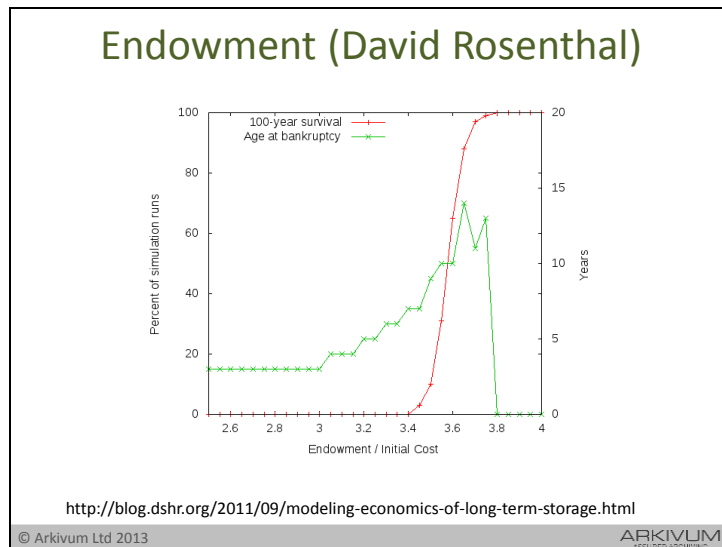
Much better is a model developed by Stephen Abrams at the California Digital Library that includes a wider range of activities and whether they are one off or recurring costs. This is based on a model of services and workflows.

<https://wiki.ucop.edu/display/Curation/Cost+Modeling>



They've developed a discounted cash flow model for the long term cost. DCF is just an accounting term for modelling the ongoing rise or fall in the cost of something over time, be it interest accumulated in a savings account or the falling cost of buying new storage each year.

His model includes the effect of earning interest on money in a fully paid-up model, i.e. where a fixed sum is paid up front to secure long-term storage. The interest accrued can be used to cover some of the ongoing costs and hence the initial sum needed is actually lower than expected by the Princeton model.

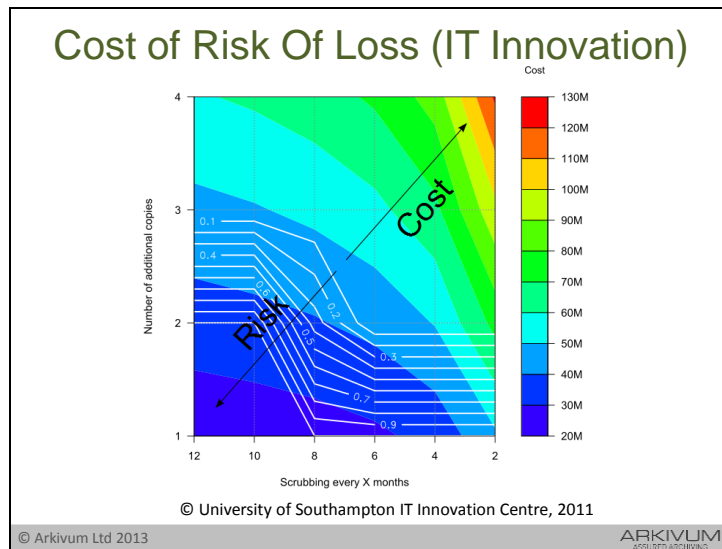


But the problem with the Princeton and CDL model is that they don't take into account uncertainty and variations.

This might be fluctuations in interest rates or fluctuations in storage costs – and the Thailand floods wiping out some major hard drive manufacturing plants is a great example of the latter. Disk costs went up after the flood and took over 2 years to get back to where they were before, whereas if they'd followed the normal trend then the costs would have halved by then. This has the effect of increasing the long-term cost by up to 50%.

David Rosenthal from Stanford has developed stochastic simulations that model these fluctuations to look at a range of possible futures. So, for example, he's modelled the probability of a given endowment being enough to cover the long-term costs of 100 years of storage.

<http://blog.dshr.org/>

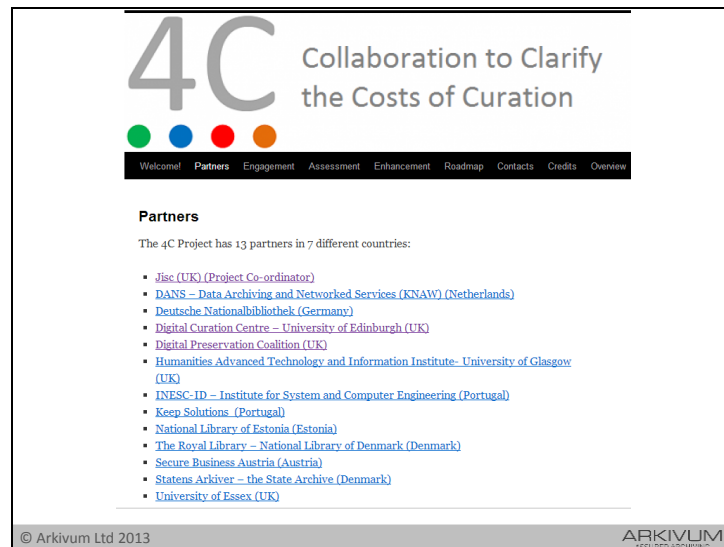


I did likewise in my previous work at Southampton, but in my case looked at the risk of loss of data over time because storage systems and the people that run them tend to fail in one way or another.

This means the need to make multiple copies of data and check and repair these copies to make sure no data loss occurs. But how many copies, how often to check, and at what cost, and with what chance of data surviving intact?

Again this is about using a stochastic simulation approach. As you invest more in storage and make more copies, or check them more regularly, then the risk of loss goes down, but the cost goes up.

http://www.rassc.org/publications/RASSC_D2.2_v1.pdf



It should be clear by now that there's a lot of work out there on cost modelling. It's getting increasingly sophisticated. It uses stochastic techniques and accounting practices such as DCF and NPV. But this plethora of tools are becoming ever harder to use and aren't embedded in the wider digital preservation process, or in standard accounting tools either, which is a problem.

This is one of the reasons why there's a new European project called 4C. The aims of the project are roughly speaking four fold:

- Understand what the requirements are for cost models across a range of sectors, which includes research data
- identify the solutions that already exist and where there are gaps
- pull together and make all the cost modelling stuff out there easier to use.
- Establish a curation cost exchange for the community to find and share cost information

I'm one of the members of their advisory board and was at a meeting with them yesterday. The project looks really good and is focussed on doing useful stuff for the community, so looks to be the definitive place to go for cost data. I'd encourage everyone to look at what the project is doing and get involved. There's also a call out at the moment for people to participate in their stakeholder analysis.

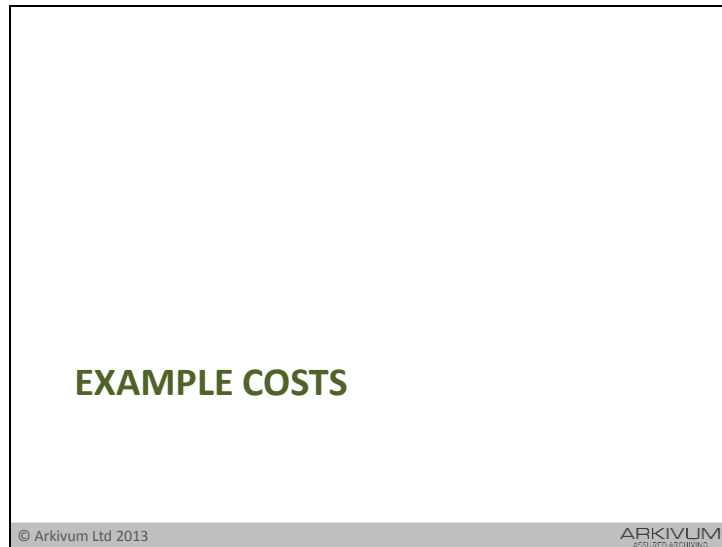
<http://4cproject.net/overview/>

Summary

- Many different cost models and techniques
- Long-term costs are hard to predict
- Risk based models are emerging
- Easy to miss out key cost elements

So, in summary, there's lots out there on cost modelling, but it's not an easy job, especially when trying to forecast the long-term.

It's also very easy to miss out important cost components, for example recurring costs based on having to deal with constant updates and migrations, which includes the cost of running pilots, trials and test systems before things go into production.



OK, so lets move on to look at some examples of what the actual costs are.



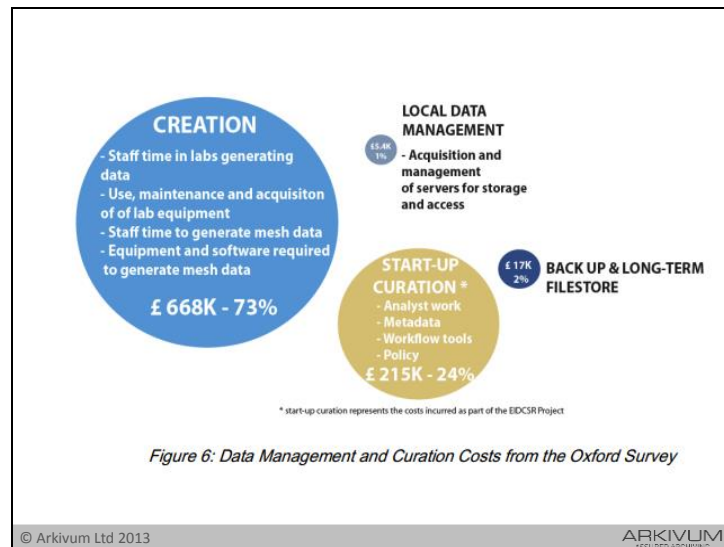
So coming back to that KRDS project. They looked at the costs incurred at various archives and institutions according to that framework I showed earlier.

Summary of KRDS2 Data Survey Responses								
Collection	Repository Type		Cost Information					
	Research	Cultural Heritage	Pre-archive	Archive	Access	Support Services	Dates	Accessible?
UK Collections								
ADS	•		•	•	•	•	2004 - Present	Possibly
BADC	•		•	•	•	•	2001 - 2008	Possibly
eCrystals	•		•	•			2002 - 2009	Possibly
EDINA	•	•	•	•		•	2006 - Present	Possibly
Linnean Soc	•		•	•	•	•	2007 - Present	Possibly
NDAD		•	•	•	•	•	1997 - Present	Possibly
NLW		•	•	•		•	2007 - Present	Yes
Oxford	•		•	•		•	2007 - 2009	Possibly
Rutherford	•	•	•	•	•	•		Possibly
UKDA	•			•	•	•	2009	Possibly
VADS	•			•	•	•	2008	Possibly
International Collections								
BABS	•	•	•	•				No
DANS	•		•	•	•	•	2008	Possibly

Figure 1: Summary of KRDS2 Data Survey Responses

The survey covered quite a wide range of types of content, from archaeology through to crystals and social science data.

Much of it was research data which makes the results particularly interesting.

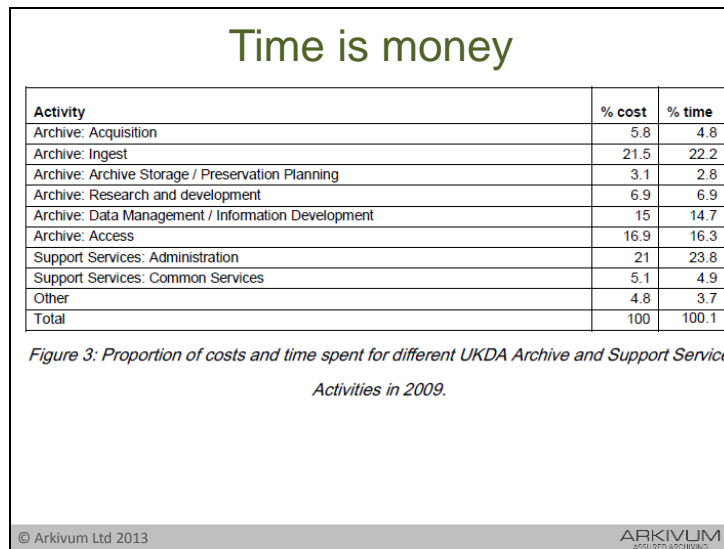


There isn't time to go into detail, so I'll pick out a couple of examples.

The first is from Oxford. The size of the circles represents the costs for each stage.

There are two take aways here

- The total cost of curation, data management and long-term storage are a lot less than the cost of the research that created the data. This means that RDM is effectively a marginal cost of research.
- The long-term costs are, in this case, a lot smaller than the initial curation costs, which means RDM costs are hugely front loaded.

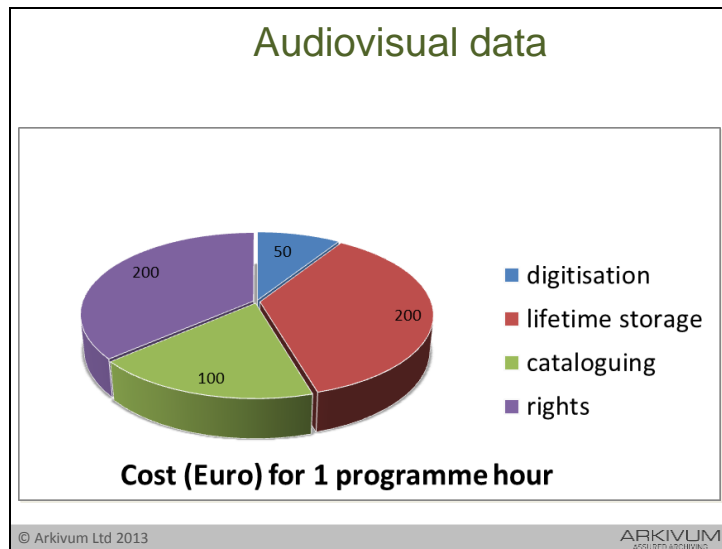


The data from the UK Data Archive for social sciences data paints a similar picture.

The case study includes the people costs, which are the dominant factor. There are over 50 people at the UK Data Archive and the total data being stored is only a few TB.

The costs again are hugely front loaded and the long-term costs are small.

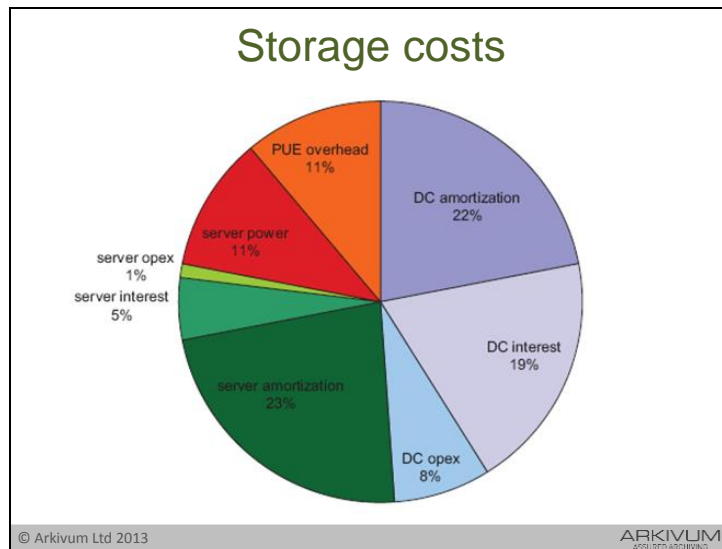
This shows that the first place to drive costs down is curation as this dominates. An interesting approach here is to question whether the cost of doing very careful selection and curation is so high that it's actually cheaper just to store everything and worry about selection later.



But it's not always the case that the storage costs are not substantial. This is some work I did last year looking at the costs of audiovisual preservation by broadcasters and national archives.

Things like cataloguing, i.e. metadata, along with rights management are the dominant factors. But because AV files are so large, storage is significant – and also unavoidable.

This is true for research too. Audio, images and video content is increasingly generated in the social sciences as well as by medical, environment and engineering applications. Other types of big data include gene sequencing, MRI and CT scans, lab equipment, and environment sensors to name but a few.



And storing this stuff is costly because of the data volumes and need to ensure it is safe for decade level retention periods.

The costs are of course a lot more than just the media and kit, but include a set of overheads including power, space, cooling, procurement, maintenance, migrations, management and more.

This chart is from Google showing the cost breakdown in their data centres.

TCO over time: storage services

- San Diego Supercomputing Centre (SDSC)
 - \$1500 per TB per year on disk in early 2007
 - \$1000 in 2008
 - \$650 end 2009
 - \$390 mid 2012
- Amazon S3
 - \$1800 per TB per year in early 2007
 - \$1260 end 2009 (over 500TB)
 - \$950 end 2010 (over 500TB)
 - \$660 start 2013 (over 5PB) **but** \$1100 for first TB
- Prices halve every 3-4 years

© Arkivum Ltd 2013ARKIVUM
POLYMER STORAGE

Looking at storage service providers like Google is one way to get a handle on the true TCO for storage.

By using the Internet Archive's Wayback machine it's also possible to look back on how the prices have fallen over time.

David Rosenthal says that the rate of fall is too low and that cloud providers are keeping costs artificially high or aren't competitive with in-house solutions. But this doesn't take into account that as volumes go up, prices go down, and looking at the entry level cost for 1 TB is a bit misleading. For a substantial volume of data, the price is falling, although not as fast as used in the Princeton model.

<https://cloud.sdsc.edu/hp/index.php>

<http://aws.amazon.com/s3/pricing/>



DuraCloud Subscription Plans - [Get Plan Information Now!](#)

	DuraCloud Preservation Basic ¹	DuraCloud Preservation Plus ¹	DuraCloud Enterprise ¹	
			Standard	Premium
Price (all bandwidth and compute charges included) this price includes	\$1,500/year for first TB \$1,000/year for additional TBs SUBSCRIBE	\$2,500/year for first TB \$1,700/year for additional TBs SUBSCRIBE	\$5,900/year for first TB \$1,000/year for additional TBs SUBSCRIBE	\$6,900/year for first TB \$1,700/year for additional TBs SUBSCRIBE
Number of redundant copies ¹	2	4	2	4
Number of cloud data centers storing content ¹	1 (Amazon)	2 (Amazon and S3SC)	1 (Amazon)	2 (Amazon and S3SC)

© Arkivum Ltd 2013 ARKIVUM
PIONEER IN CLOUD ARCHIVAL

And then there are specific preservation services that sit on top of cloud providers, with DuraCloud being an example in the US. They manage replication and integrity across a set of storage locations to provide more assurance of long-term data safety, but this of course comes at an additional cost. Their service is \$1700 per TB per year if you want content spread across more than one cloud storage provider – and that’s after you’ve got over the hurdle of the higher costs of the first TB.

<http://duracloud.org/content/pricing>

Internal storage services			
Institution	Cost (£ per TB)	Duration (years)	Approach
Bristol	1500	20	One disk copy, tape backup
Southampton	1000		One disk copy, tape backup
Essex	1459		One disk copy, tape backup
Bath	850	15	Offline tape after 5 years
East Anglia	950	5	One disk copy, tape backup
Oxford	4200	5	HSM, three copies on tape

- £450 per TB per year single copy
- £5000 per TB POSF multiple copies

© Arkivum Ltd 2013

ARKIVUM
POLICY CENTRE

There are also several Universities that run their own internal storage services. These are some examples where the prices have been published on the Internet. The prices vary widely depending on how the data is stored, how safe it is, and how long it will be kept for.

At a recent JISC workshop attended by a wide range of Universities to discuss storage services, the spectrum of costs was between £450 per TB for a single copy for 1 year through to £5000 per TB fully paid-up for indefinite storage using multiple geographically dispersed copies of the data.

<https://www.acrc.bris.ac.uk/acrc/storage.htm>
http://www.southampton.ac.uk/library/research/researchdata/storage_options.html
<http://www.sussex.ac.uk/its/services/research/researchdata>
http://www.bath.ac.uk/bucs/aboutbucs/policies-guidelines/research_data_storage_guidelines.html
<http://www.uea.ac.uk/is/storage/cost-calculator>
<http://www.oucs.ox.ac.uk/internal/sld/hfs.xml>



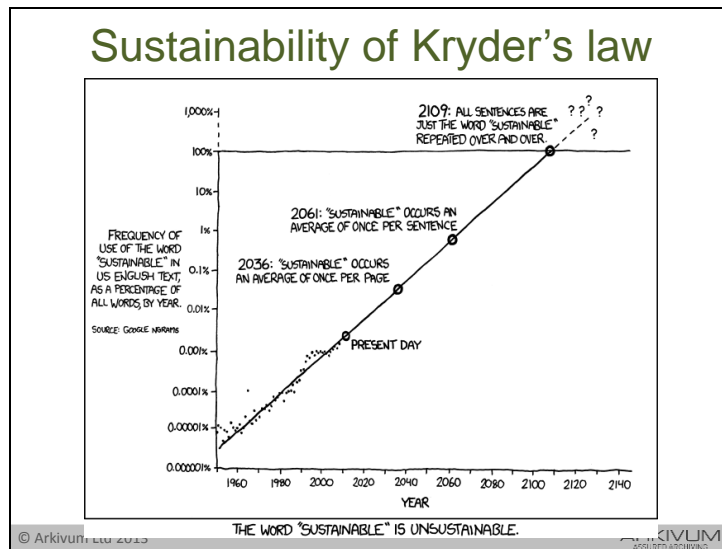
Many Universities have based their costs on the Princeton model, i.e. estimate current costs and then apply a multiplier to calculate the lifetime costs.

This is fine provided that you know the trends for future costs.

This is a magazine advert from 1980. A 10 MB hard drive for \$3k. Today you can get 10 TB for close to \$300. That's a 10 million increase in capacity for the same cost.

This ongoing increase in capacity for the same money is called Kryder's Law, which is just the storage equivalent of Moore's law.

http://en.wikipedia.org/wiki/Mark_Kryder

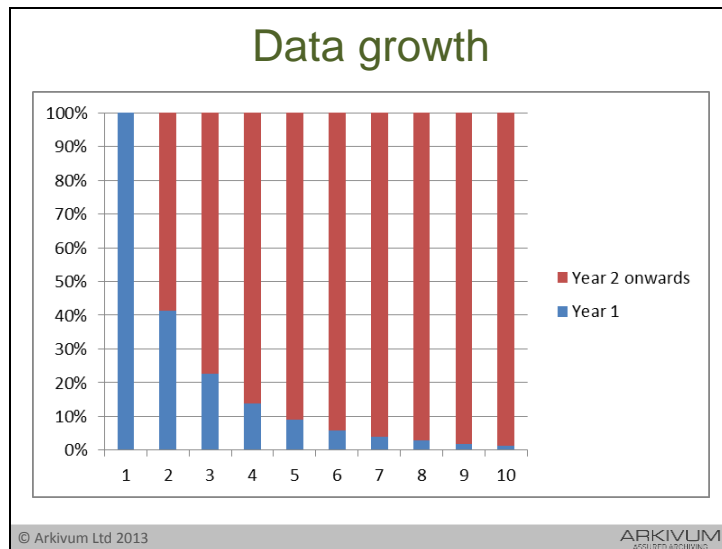


It's tempting to assume that Kryder's law can be maintained

But as this amusing chart shows, all good things come to an end eventually. The diagram shows the risks of trying to extrapolate too far into the future based on past performance

In this case, its about the use of the word 'sustainable' in sentences. Looking at the increase in usage from the past and projecting it forward would soon imply that every word in every sentence will be the word sustainable.

There's increasing evidence that Kryder's law is slowing down, especially for disc storage, so this needs factoring to long term costs.



But there are other bigger factors that influence cost.

The first is data volume.

This chart shows an archive that is doubling in data volume every 2 years. The blue bar is the fraction of the archive that is occupied by data ingested in year1.

The blue part very quickly becomes a small part of the total data being stored.




The data in year1 has become a marginal part of the archive after just 5 years.

Therefore, the cost of storing this data is a marginal cost compared to new data in the archive.

So, calculating future costs of storing data has to be based on marginal costs in the context of how much new data needs to be stored too.

This is something that most calculations, e.g. Princeton, miss out.

Cost per TB falls as volume increases

• £130 per TB	TS3100	
• £100 per TB	TS3310	
• £47 per TB	TS3500	

© Arkivum Ltd 2013

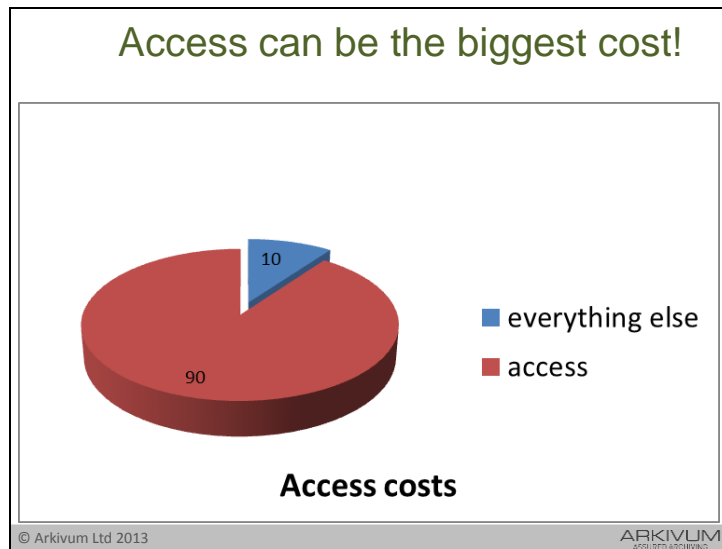
ARKIVUM
Tape Libraries

And it's important not to miss it out because the cost of storing data is very much dependent on how much you want to store.

There are big economies of scale to be had.

This table shows a trivial example of different size tape libraries. As the libraries reach scale, the cost per TB falls significantly.

The point is that unless data volumes in an archive keep growing then it's hard to maintain these economies of scale into the future, which means long-term costs can be higher than expected.



Next comes access costs. These can dwarf all other costs. This chart goes back to that audiovisual example I had earlier and shows the relative size of access costs when content is made available online for free access. The network costs for the broadcaster concerned were alone bigger than all the other costs combined.

Access is of course a concern too for research data, especially with a drive to open access as pushed by the research councils. This leads to the worry that future unquantifiable access costs will have to be born by institutions and there is no way of recovering them from the users, i.e. its an 'author pays' model.

Egress costs

- Amazon Glacier
 - Storage \$0.011 per GB per month
 - Ingest Free
 - Retrieval Free 5% per month*
 - Internet Access \$0.012 per GB

* Pro-rata on daily basis.

- Retrieval can be 100x monthly storage charge

© Arkivum Ltd 2013ARKIVUM

And the access cost worry is amplified when considering some services for long-term storage, for example Amazon Glacier.

Their storage costs are very attractive, but when it comes to accessing the data the costs can rocket. Glacier allows 5% of data to be retrieved each month for free, but this is pro-rata on a daily basis. If you retrieve more than 0.16% on a given day, then you can be charged a whole month's worth of excess retrieval charges, which can be 100x higher than the storage cost at that point.

<http://aws.amazon.com/glacier/#pricing>

Cost recovery - funders

Principle 7 – the costs of data management are indeed payable, since ‘It is appropriate to use public funds to support the management and sharing of publicly-funded research data’. The aim of the councils is to maximise research benefit. That means using money efficiently.

Principle 2 – ‘Data with acknowledged long-term value should be preserved and remain accessible and usable for future research.’ Consequently, data management plans should exist, both at an institutional level and at a project level so that data with acknowledged long term value can be made accessible and reusable..”

© Arkivum Ltd 2013

ARKIVUM
RESEARCH COUNCILS

<http://www.dcc.ac.uk/blog/conversation-funders>

Now comes the question of recovering the costs of RDM.

JISC and the DCC organised an event at Aston University on 25th April where all the major UK funding bodies were present to answer questions on RDM costs and what was eligible on a grant.

The good news is that RDM costs are eligible in one way or another.

The research councils fund research in the expectation that it produces a public good that should be made accessible. Data should be produced from good science and should be capable of being used to verify that science. This means that good RDM is an inherent part of what the funders will pay for as it is an integral part of good scientific practice.

This is seen in the RCUK principles of data management being an eligible cost and that data with value should be retained and be made accessible – which includes primary data that underpins research publications.

Direct or indirect

- Direct costs: preparation, metadata, deposit
- Indirect cost: long term storage, preservation
- Can't double charge
- Don't need to keep everything – selection

Good science is repeatable and verifiable
Public money creates a public good

© Arkivum Ltd 2013

ARKIVUM
DIGITAL CURATION CENTRE

Some costs are eligible as direct costs on a grant, e.g. preparing a specific data set in a project so it can be kept, either internally or by depositing in a national repository.

Some activities, e.g. long-term storage, require the use of infrastructure that is shared within an institution, so is not specific to a given project. Generally speaking, the costs of this are included in an institutions overheads and hence can be recovered through indirect costs on a grant – getting the money together for the initial capex investment is a different matter – its more a case of 'spend now, recover later'

The Digital Curation Centre provides a good review of funder requirements, so that's a good port of call for more information.

And they've also summarised the workshop with the funding bodies

<http://www.dcc.ac.uk/blog/conversation-funders>

Summary

- Lifecycle costs are highly variable
 - Type of data and community
 - Curation
 - Access
- Long-term retention needs to be calculated on a marginal cost basis
 - Data volumes
 - Realistic trends
 - Factor in uncertainty

© Arkivum Ltd 2013

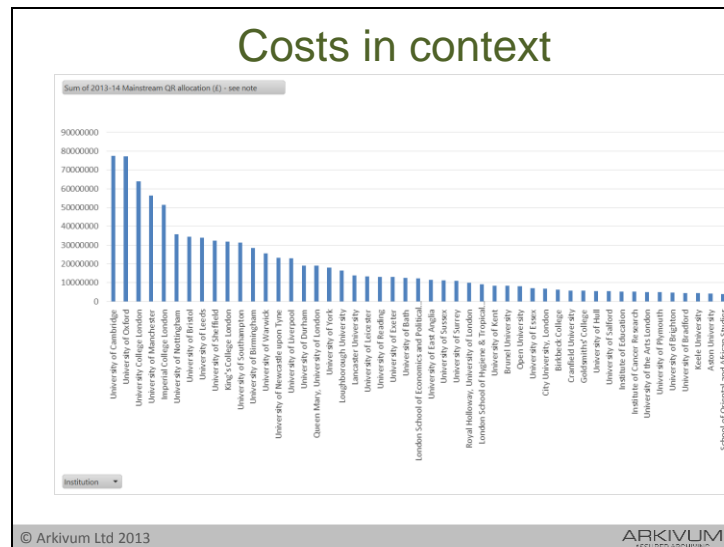
ARKIVUM
POLICY CENTRE

So to summarise:

The costs are variable, can be very much front loaded, are substantial when a lot of curation is involved, and are hard to quantify in many areas, e.g. access.

Long-term costs for data retention need to be worked out on the basis of data growth, realistic trends in storage costs, and take into account a good measure of uncertainty, i.e. they need to include contingencies.

Costs can be recoverable from funding bodies, but the model to do this depends on the type of RDM activity involved and to some extent the policy of the particular funding body.

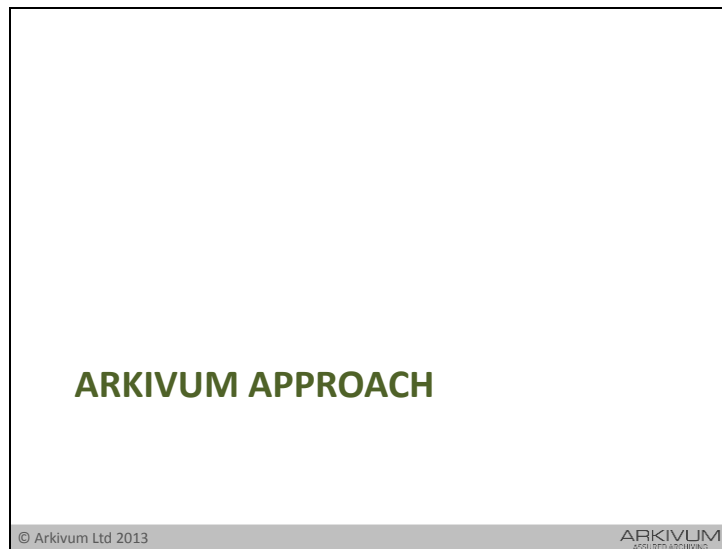


Then finally its worth putting RDM costs in context.

Compared with the budget for the original research that generated the data or the QR funding from HEFCE for high quality and high impact research, the cost of long-term retention and access to research data is small.

For example, the research income of the larger Universities is measured in 100s of Millions of £. The QR money they get from HEFCE will be in the 10s of millions or more. The total cost of RDM is typically a few million. i.e. RDM is a small % of the total budget that Universities have and really is a marginal cost.

Slide 45



OK, so now for the last few minutes of the webinar where we'll look at Arkivum's service.

A presentation slide for Arkivum. At the top, the word "Arkivum" is written in a green, sans-serif font. Below it is a bulleted list of five points: "Online data archiving as a service", "Spin-out of University of Southampton", "Decade of know-how working with archives", "Safe, secure, accessible data storage", and "Designed from ground-up for retention and access". Below the list is a dark grey rectangular box containing the Arkivum logo on the left and a "100% data integrity guarantee" with the tagline "Keep your data safe & secure forever" on the right. The logo consists of the word "ARKIVUM" in blue and yellow, with "ASSURED ARCHIVING" in smaller blue text below it. The footer of the slide is a light grey bar with "© Arkivum Ltd 2013" on the left and the Arkivum logo on the right.

Arkivum

- Online data archiving as a service
- Spin-out of University of Southampton
- Decade of know-how working with archives
- Safe, secure, accessible data storage
- Designed from ground-up for retention and access

ARKIVUM
ASSURED ARCHIVING

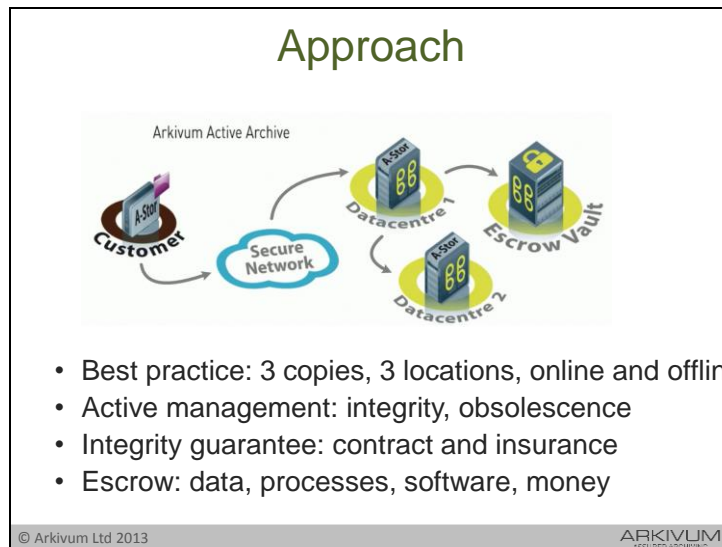
100% data integrity guarantee
Keep your data safe & secure forever

© Arkivum Ltd 2013

ARKIVUM
POLICY CENTRAL

Arkivum provides online data archiving as a service for those organisations that need to keep data for the long-term for compliance or reuse. We have customers in construction, life sciences, energy and voice call recording to name but a few.

The company was founded in 2011 as a spin out of the University of Southampton and is based on expertise the founders have from working with large archives, for example national broadcasters and archives, on how to retain digital assets for the long-term.




The approach we take is to follow data preservation best practice. Three copies of customer data are held in three locations. We use checksums to actively manage data integrity through regular checks and we do regular media and infrastructure migrations to counter obsolescence and to ensure costs remain low. Basically, we take care of that 'blue and yellow stuff' in that diagram I showed earlier. And it's our skilled and dedicated staff that do this that you're also getting access to, not just the infrastructure we use. Good preservation practice, trained staff and very carefully controlled processes means we can offer a guarantee of data integrity.

All data returned from our service is always bit-for-bit identical to the data the customer supplied, with no restrictions on time or volume. The guarantee is backed by insurance and is included in our SLA. We are also certified to ISO27001 and Arkivum has been externally audited for the integrity, confidentiality and availability of data assets in our possession.

Finally, one of those copies of customer data is held at a third-party site with a three way agreement in place with us, the third-party and our customers so that if they want to leave our service, or if we can no longer provide the service we agreed, then the customer gets direct access to a complete copy of their data on media that they can take away. This is part of a wider escrow model that includes the software and processes we use to run the service as well as ring-fencing of money for fully paid up long-term retention contracts.

Pricing

- PAYG or Paid-Up for 5,10 or 25 year
- JANET, no ingress or egress charges
- Escrow included, no exit cost
- ISO27001 included, external audits built in



© Arkivum Ltd 2013 ARKIVUM
POLICY CENTRE

Which brings me on to pricing.

We support a PAYG model, but more interesting to the research community is our ability to offer fixed-term contracts that are fully paid-up, i.e. a capex model for a service. Offering paid-up long-term contracts means we fit well with cost recovery from research grants or capital budgets within University IT services and departments.

We'll be on JANET by the end of the quarter and we don't charge for ingress or egress of data.

For comparison, our prices are lower than Amazon S3 or other enterprise storage services, lower than DuraCloud, and competitive to the POSF costs that institutions are currently calculating for an equivalent in-house approach.

So, an example, is that educational list price is £1500 per TB paid-up for 5 years of storage. After that the price can be fixed, e.g. at 50% or less for renewal. Or you could go for a longer-term contract at the offset.

Conclusions

*"Digital information lasts forever -
or five years, whichever comes first."*
Jeff Rothenberg

Message: Long-term RDM is not cheap!

Arkivum: helps keep the costs manageable

© Arkivum Ltd 2013

ARKIVUM
RESEARCH DATA MANAGEMENT

Just to re-iterate the original message: RDM can be expensive - but the costs are often still a small fraction of doing the research in the first place, and the costs can be recovered from funding bodies or other sources.

I hope the webinar has been helpful in how to calculate how expensive and what the actual costs might be.

Arkivum can fulfil the long-term data retention and retrieval aspect of digital preservation.

We might not be the cheapest way to store data, but instead we do offer guaranteed data safety as part of our contract and we are competitive with in-house alternatives.

The costs can be fixed and hence quantified in advance, including no cost for data egress or exit from our service. This makes it a lot easier for institutions to predict, manage and recover their costs, e.g. from funding bodies.

Questions?

www.arkivum.com
matthew.addis@arkivum.com

"Security is of key importance to our business, Arkivum's A-Stor service allows us to store our encrypted data for the long term in a cost efficient way that is entirely scalable and reduces pressure on our internal IT infrastructure." Dan Watkins, Oxford Fertility Unit

© Arkivum Ltd 2013

ARKIVUM
POLICY CENTRAL

One final thing – we’re doing a kind of tour round Universities at the moment where we are happy to call in and discuss openly all aspects of research data retention and access, including our experience of what other Universities are doing, what the funding bodies are trying to achieve, and how we can fit into the picture. There’s more details on our website, but if you are interested in us paying you a visit to cover anything in this webinar or the ones before it then just let us know and we’d be happy to oblige.