



Good afternoon. I'm Matthew Addis and CTO of Arkivum.

This webinar is the first of three webinars that we have planned over the next month or two.

They are essentially a re-run of webinars on Research Data Management that we ran earlier this year, but I have updated them in places as RDM is a fast moving area and quite a lot has happened in the community over the last few months, including some very interesting events organised by JISC and the DCC on topics relevant to these webinars, e.g. storage as a service and what funders consider to be eligible costs for RDM on a grant application.

And by way of background, previous to joining Arkivum full time at the start of the year, I spent 16 years at the University of Southampton IT Innovation Centre leading collaborative R&D projects, including digital preservation projects in sectors such as audio visual and aerospace where I've worked with national archives and large industrial organisations on long-term retention and use of digital assets. This means that as well as being experienced in digital preservation and archiving, I've also been on the 'research' side of the fence and know what it's like to work inside a University.

Webinar series

- 1 Effective Long-term retention and access of Research Data
Matthew Addis
- 2 Long-term data management: in-house or outsource?
Neil Beagrie
Matthew Addis
- 3 The real costs of long-term data retention
Matthew Addis

© Arkivum Ltd 2013

ARKIVUM

1. Today's webinar is an overview of landscape and issues, including how Arkivum's solutions can help.
2. Next is an exploration of the pros and cons of in-house or outsourcing all or part of data retention and access, where we've included a great contribution from long-standing expert in preservation and RDM Neil Beagrie, who's provided a vendor independent viewpoint and given his expertise on the question of shared services and cloud models.
3. In the third webinar we explore the real through-life costs of data retention and access, including what makes up the TCO and strategies for reducing, sharing and crucially how to recover those costs from funding bodies or other sources.

Contents

- The drivers for keeping research data
- Requirements from the funding bodies
- Decade level data safety and security
- Breaking the problem down
- Arkivum's solutions

"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 Clinic

© Arkivum Ltd 2013 ARKIVUM

This webinar is split roughly into two halves.

In the first half we look at what's driving institutions to keep their research data, what the funding bodies are asking for, and how retention and access are actually flip sides of the same coin. This will be quite a broad overview and exploration of what is a complex and rapidly evolving landscape – I'll cover some of the many issues including policy, culture change, what's happening at the institutional level, and some of the specific technical challenges around metadata, formats, storage and access.

In the second half, we'll get more deeply into the issue of long-term data retention, the challenges of decade level data safety and security, the interplay between safe storage and easy access, and then I'll wrap up by spending the last part of the webinar looking at how Arkivum's solutions can help.

This will take about 30 minutes and after that we'll be open for Q&A.

I'll go quite quickly and there's quite a lot of information, but the recording of this webinar will be available afterwards along with slides and full set of speakers notes which includes URLs to all the sources of information that I'll reference.

Why keep research data?

- Funding body requirement
- Protection of IP
- Regulatory compliance
- Sharing, reuse
- Save money through archiving

© Arkivum Ltd 2013

ARKIVUM

So why is there a need to keep research data? There are a multitude of reasons. For example,

The bodies funding the research mandate that the data has to be kept and importantly be made accessible, for example the EPSRC stipulate retention for at least 10 years beyond the point that a given bit of research data was last accessed. This is all about encouraging institutions to maximise the public value of the research they do and to benefit from doing this rather than an intention to become a regulator with a big stick and enforce compliance – but more on this in a minute.

Moving down the list of reasons to keep data, it might be that it supports a patent application, i.e. there are intellectual property drivers, which are increasingly important as Universities seek to exploit their research as part of their enterprise agendas.

There are compliance issues, in addition to funding body requirements, e.g. data relates might relate to a FOI request, or it's personal data and hence falls under DPA, or because the data supports something like a clinical trial or drug development where legal requirements from GCP or GLP apply. Here it all gets a bit messy as there can be conflicts between making data accessible, keeping data because an access request has been instigated, withholding data access because of confidentiality and privacy, having to delete data because it is no longer being used for the purpose it was originally collected, and having to keep data because of its relationship to something outside of the activity of those who generated the data, e.g. lifetime of a drug compared to a clinical trial on some specific aspect of drug efficacy. The 'climategate' [1] issue of the CRU at East Anglia is a good example of the hot water that it is easy to get into where hacking released emails that triggered FOI requests and review of validity of existing research.

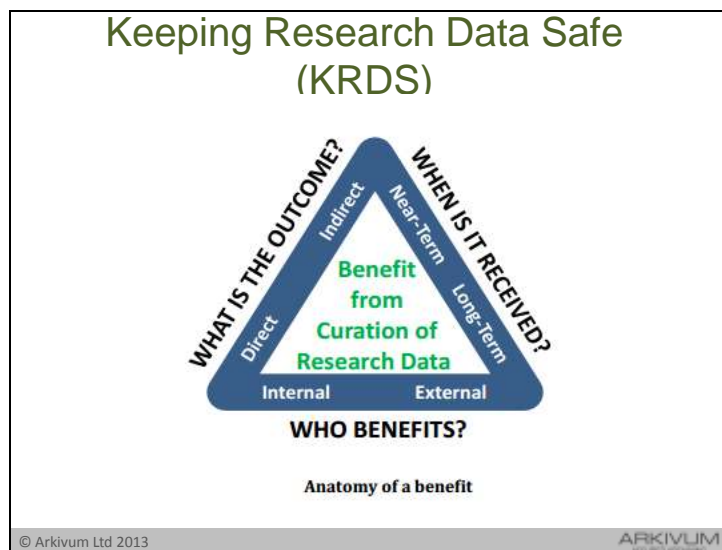
There is value in sharing or reusing data, e.g. new research based on historical data, or because publishing the results of research in a form that allows others to use the data increases the impact of the original research, which in turn leads to prestige and reputation of the researchers and institution and further collaboration. This is a major positive benefit of keeping data. Evidence of data sharing and reuse can then be used as part of Research Assessment to generate more income (REF2014 already has more emphasis on impact of research than then RAE and this will surely only increase) or it might lead to more research funding – or perhaps this is better phrased as helping a University being more competitive when it comes to securing funding compared to other Universities. Retention for sharing and reuse has the potential to be a competitive edge with very tangible benefits. A large research led University can easily bring in £100M of research funding each year. A few percent increase in this through good RDM has £M benefits and more than outweighs the investment needed.

Finally, although not a driver to keep data per se, moving data that does need to be kept to archive can have the valuable effect of saving money. The systems used for holding data when it is first created and processed are really expensive places to store data for the long term. Moving data to a proper archive system can save money, not just by reducing primary storage needs but also in turn by needing less resource for associated maintenance and backup.

So overall, there are many drivers, most revolving around the need to retain data because it has future value, which is the main reason funders require retention, i.e. because of the benefits of being able to access it in the future.

[1]

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




The benefits of retaining and making research data accessible are both positive, e.g. reputation or getting more research funding, but also include the benefits of preventing future costs e.g. recreating lost data, not servicing access requests, not meeting funding requirements, or not being as successful in getting new funding.

The Keeping Research Data Safe project provides lots of guidance on the benefits of keeping research data and how to describe the value of doing so, even when intangible, which is where I've taken this picture from.

So if you are in the game of creating a business case for investment into RDM then KRDS is a good place to start.

Funder Requirements

- Retention
- Access
- Deposit with national services
- Data Management Plan
- Roadmap



© Arkivum Ltd 2013 ARKIVUM

If we look at the requirements of the funding bodies in more detail, then generally speaking they can be classified as specifying a minimum retention period, the need to provide some form of access, sometimes a requirement to lodge a copy of the data with a national repository (if one exists) and in almost all cases to provide a data management plan of how this will be achieved, which needs to be part of the grant application.

At a meeting organised by the DCC on the 25th April, representatives were present from all the major funding bodies to answer questions on what aspects of RDM they would fund through grants. There was a clear message that the reason for mandating retention of research data, requiring open access unless there is a good reason not to, and needing data management plans with justified costs is all because the research councils see value in research data being made accessible and reused – it's all about the public good of making research data funded from public money available. The requirements for doing this in a properly planned and managed way is just a reflection of the expectation for Universities to have the necessary infrastructure and good research practice in place to deliver high quality and sustainable research output. This is really important as the emphasis is on encouraging Universities to maximise the value of the research they do and to adopt good practice as a matter of course – its not a desire of the funding bodies to become policemen and try to monitor and enforce research data retention.

But implementing this isn't the work of moments for institutions, and the funders have set timescales for having this all in place, for example 2015, and in some cases, for example the EPSRC, they require institutions to provide a roadmap of how the transition will be made.

The Digital Curation Centre provides a good review of funder requirements, so that's a good port of call for more information [1]

[1] <http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx>

Slide 7

Curation policies and support services of the main UK research funders											
Research Funders	Policy coverage		Policy stipulations					Support provided			
	Published outputs	Data	Time limits	Data plans	Access / sharing	Long-term retention	Monitoring	Guidance	Repository	Data centre	Costs
Arts and Humanities Research Council	●	●	●	●	●	●	○	●	○	●	○
Biotechnology & Biological Sciences Research Council	●	●	●	●	●	●	●	●	●	●	●
Cancer Research UK	●	●	●	●	●	●	●	●	●	○	○
Engineering and Physical Sciences Research Council	●	●	●	●	●	●	●	●	○	○	●
Economic and Social Research Council	●	●	●	●	●	●	●	●	●	●	●
Medical Research Council	●	●	●	●	●	●	○	●	●	○	○
Natural Environment Research Council	●	●	●	●	●	●	●	●	●	●	●
Science and Technology Facilities Council	●	●	●	●	●	●	●	●	●	●	○
Wellcome Trust	●	●	●	●	●	●	●	●	●	●	●

Terminology clarifications:
Published outputs: a policy on published outputs (e.g. journal articles and conference papers)
Data: a data policy or statement on access to and maintenance of electronic resources
Data plans: set of mechanisms for making content accessible or preserving research outputs
Data sharing: requirement to consider data retention, management or sharing in the application
Access/sharing: promotion of OA journals, deposit in repositories, data sharing or reuse
Guidance: stipulations on long-term maintenance and preservation of research outputs
Monitoring: whether compliance is monitored or action taken (such as withdrawing funds)
Guidance: provision of FAQs, best practice guides, toolkits, and support staff
Repository: provision of a repository to make published research outputs accessible
Data centre: provision of a data centre to store validated electronic resources or data
Costs: a willingness to meet publication fees and data management / sharing costs

KEY:
● full coverage
◐ partial coverage
○ no coverage

<http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies>

© Arkivum Ltd 2013

ARKIVUM

The DCC comparison of funder requirements is seen in this chart. The requirement to retain and provide access to data as well as publications is universal. There's also the recently updated RCUK guidelines on open access to publications, which includes the data they are based on, which further emphasises that access is the main driver for retention.

But the support required to do this, for example practical guidelines, availability of national repositories, or the recovery of costs from grants varies quite a lot. Here it's an interesting case of the funders specifying WHAT they want to happen, but not saying much about HOW to make it happen, or providing full support for the costs.

Whilst the many RDM projects have done some really great work, and there's plenty of further support from the DCC or preservation community in general, the one thing that is still missing is consolidated, consistent and practical guidelines on how to do RDM effectively – especially so that institutions, or which there are many, that weren't involved in the JISC MRD programme can roll-out their own RDM strategies and infrastructure. Hopefully we'll see this emerge in the next phase of RDM

Example of specific requirements (1)

- MRC GRP guide



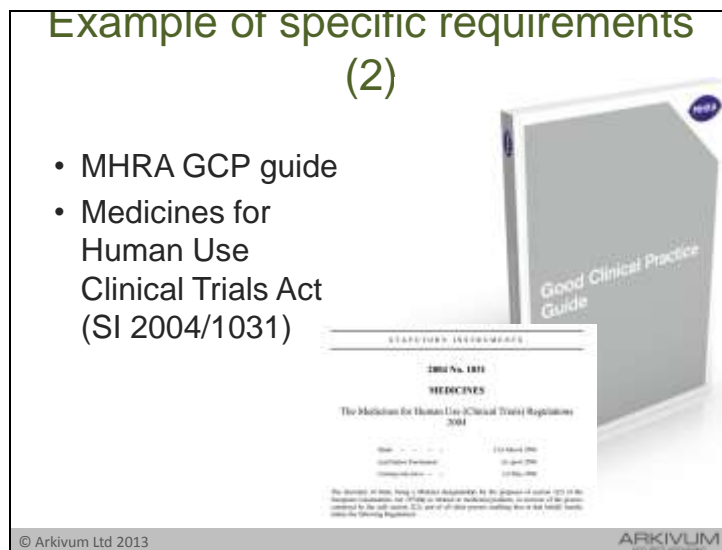
© Arkivum Ltd 2013

ARKIVUM

When it comes down to specific funding bodies and types of research, the situation can get quite complicated. So, for example, the Medical Research Council publishes guidelines on Good Research Practice [1], which includes the need to retain research data and have a data management plan.

[1]

http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/Researchpractice/principles_guidelines/index.htm




But if the research is part of a clinical trial for example, then further requirements apply, e.g. the MHRA Good Clinical Practice guidelines. These are quite extensive when it comes to archiving, some of which is underpinned by UK legislation through statutory instruments such as the clinic trials act 2004 which was amended in 2006 to include retention and access to clinical data. This is where we start to see specific guidelines emerging, e.g. the MHRA GCP guide includes electronic archiving.

Example of specific requirements (3)

Funders covered in this document

The first version of the funders list covers 17 agencies that have supported research activities performed by LSHTM during the past decade:

1. Action Medical Research (AMR)
2. Biotechnology and Biosciences Research Council (BBSRC)
3. Bill & Melinda Gates Foundation
4. Breast Cancer Campaign (BCC)
5. Cancer Research UK (CRUK)
6. Department of Health, UK (DoH)
7. Department for International Development (DfID)
8. Drugs for Neglected Diseases Initiative (DNDi)
9. Economic and Social Research Council (ESRC)
10. Engineering and Physical Sciences Research Council (EPSRC)
11. GlaxoSmithKline (GSK)
12. Medical Research Council (MRC)
13. Natural Environment Research Council (NERC)
14. Wellcome Trust
15. WHO - World Health Organization - TDR

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE 

© Arkivum Ltd 2013 ARKIVUM


There's a really interesting review [1] by The London School of Hygiene and Tropical Medicine of who funds their work and the requirements they impose. They need to comply with the needs of 17 different funders.


And that's for just one school. Now imagine how complex it gets at the Faculty or University level where IT services and the Library often become involved in trying to set up common services across the University for research data management. This isn't always appreciated by individual researchers who have an immediate problem to solve and can become frustrated at what appears to be slow progress at the University level. Likewise at the University level, it's much more efficient and cost effective to provide a single generic service for research data, but this may not meet the very specific needs of each research group which means barriers to adoption and reluctance to take up.

[1] http://researchonline.lshtm.ac.uk/208596/1/Funder_Requirements_Analysis.pdf

Data Management Plans

- Common requirement of funding bodies
- Often the responsibility of the PI
- Part of proposal assessment
- Includes long-term preservation
 - Strategy
 - Responsibilities
 - Retention policy
 - Storage/repositories
 - Costs

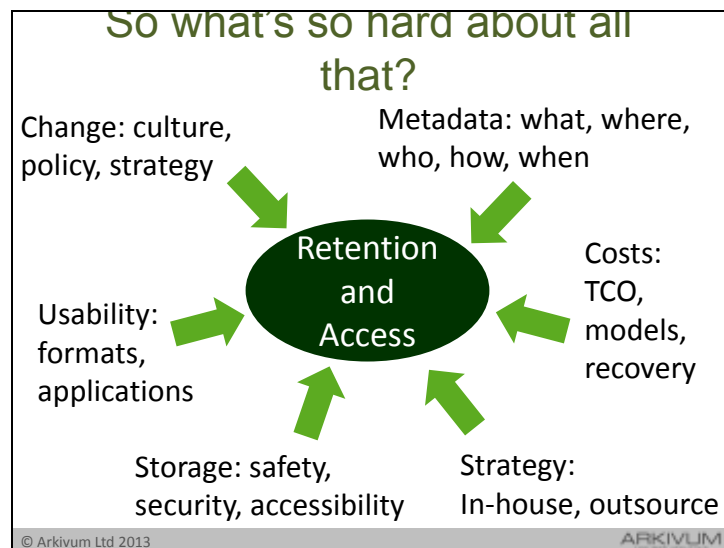


The  Data Management Planning Tool

© Arkivum Ltd 2013ARKIVUM

But whether the support for research data retention and access is provided at the University, School or individual researcher level, the need for a Data Management Plan is common across nearly all the funding bodies as part of a grant application (the EPSRC being an exception, but even they expect a case for support that shows the institution has the capacity to do the work including handle the data created). Responsibility will typically fall to the PI. The DMP needs to include long-term preservation of data with details of how it will be done in practice.

Again the DCC provides some useful guidelines and support, this time in the form of an online tool to guide people through the process of writing a DMP.



So what's so hard about RDM and keeping research data safe?

First there is often a need for change. For example, even at a basic level, there can be a need to convince researchers that storing research data on USB drives in a desk drawer simply won't cut it, which can be a challenge itself since convincing researchers to pay more for a proper infrastructure means their budget for the research may then be lower. It's at this level that libraries and central IT services try to instigate change through policies, education and piloting new services.

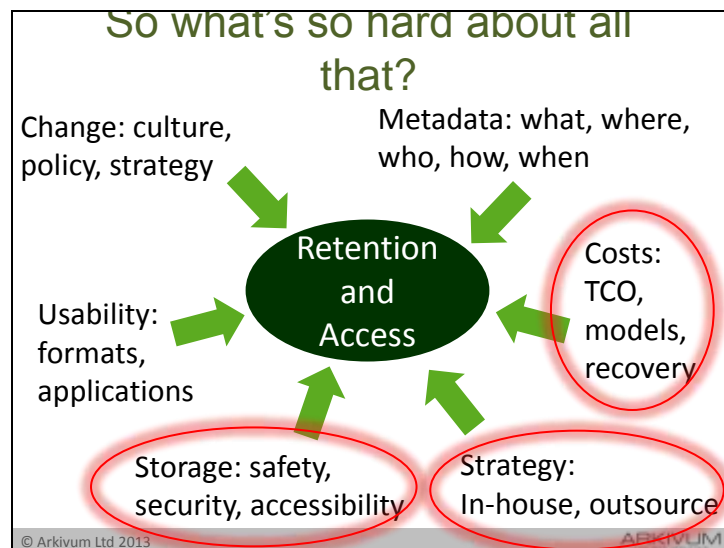
Then comes the issue of metadata. Without this there's little hope in knowing what's been held, which is the key to finding it and making it accessible, so there's the need to collect metadata on what the research data is, where it came from, who created it, what for, how and when. The challenge here, especially at the institution level, is standardisation and finding a scheme that is simple enough to be workable and adopted, but flexible enough to cope with the diversity of data types in a University setting.

But knowing what you have is no use if you can't read and understand the research data in the future, which requires appropriate data formats and sometimes the need to retain the ability to re-run the applications that created or interpreted the research data in the first place. There are a multitude of issues and techniques here including migration, emulation, virtual machines and others that warrant a webinar of its own.

None of this helps of course if the data isn't stored in a way that is safe, secure and readily accessible. This is where a plethora of local solutions chosen by researchers at an ad-hoc level quickly becomes unmaintainable and can put data at risk – but lots more on this later.

Properly handling the issues of metadata, formats, storage and access at the technical level as well as effecting change within a University all costs time and money. Budgets are under ever increasing pressure, so there are challenges of correctly calculating through-life costs and how to recover these through grants or fully account for them in overheads.

Finally, to keep these costs down, and to avoid large capital expenditure at a time where budgets are falling, especially for centralised IT, the use of shared services and opex models becomes very attractive. This then brings in the questions of how and when to outsource or to provide in-house solutions.



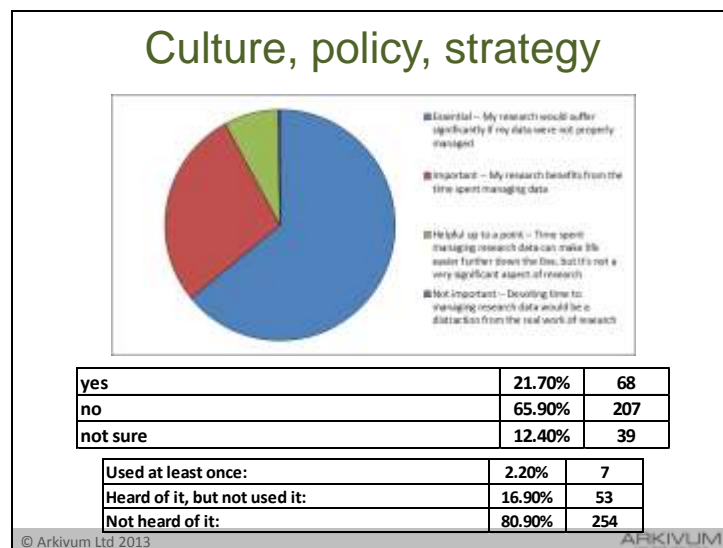
The issue of storage is the main focus of the rest of this webinar.

The next webinar will look at the pros and cons of in-house or outsourced solutions.

The third webinar will then look at the costs.

I'll touch briefly on the other topics next, but before then what I want to stress is that the challenges are manifold.

Different parts of an institution will tend to be focussed more on one area than another when looking to make inroads – for example library or central IT looking at issues of policy or metadata and researchers or schools more concerned with specific solutions, like running out of storage or short term data sharing and data reuse.



Some of the RDM projects have released the results of surveys on the need for, and awareness of, RDM - notably the DaMaRo project at Oxford, where the statistics on this slide are from, but also a more recent report from the London School of Hygiene and Tropical Medicine that I touch on in a minute.

The DaMaRo report seems to be typical – there is often a high awareness of the need to keep research data, shown here in the pie chart, but much less awareness of the services available to do so – even though there may be institutional policies in place for research data management and institutional services to help store and manage that research data. So, in the DaMaRo survey, only 21% of those surveyed knew that there was a University policy in place and less than 20% of respondents knew of, or had used, the Universities web pages on RDM.

This illustrates what many Universities find - that awareness is a major problem.

Then comes the issues of training – further survey work at Oxford showed a strong need and appetite for training, but that most training for data management was ‘on the job’ rather than formal. Furthermore, the areas where least training took place was on IP issues and preparation of data for digital preservation, which are both core to RDM.

<http://blogs.oucs.ox.ac.uk/damaro/2013/01/03/university-of-oxford-research-data-management-survey-2012-the-results/>

<http://blogs.oucs.ox.ac.uk/damaro/2012/11/21/damaro-survey-results-research-data-management-training-for-the-sciences/>



The lack of awareness, lack of training, and general ‘don’t know any better’ problem of research data management at the researcher and PI level causes all sorts of problems. Not least is the widespread adoption of approaches that can’t be considered good practice, for example storing research data on hard drives in desk drawers – as shown here in the results of a survey by the London School of Hygiene and Tropical Medicine – where there were as many, if not more, respondents storing data on local servers and portable media as there are were using centralised storage services – and that’s for an institution where much of the research is more heavily regulated than normal within Universities, e.g. because it’s laboratory or clinical work.

But it’s not just that data is being stored using inappropriate means, or is distributed across research teams and departments, or is hard to find – it’s that this sets expectations on what the cost/ease of use should be for a centralised alternative.

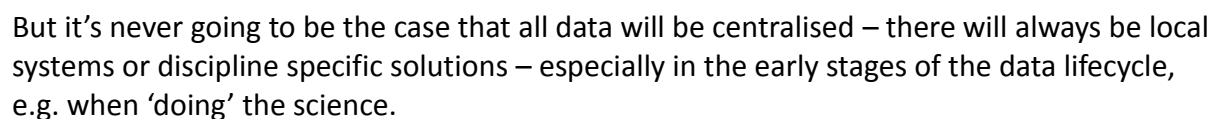
If a researcher can go down PC world and spend £50 for a TB of storage and plug that straight into their local systems to store data, then they will question the cost and benefit of being charged 10x that much to store data on a central server that can be harder to access – either from their local work environment or when using mobile devices.

This creates a real hurdle for centralised RDM services – getting hold of the data and encouraging adoption.

This is why several institutions provide storage for ‘free’ e.g. you get 1TB on a central store for no cost. Institutions are also investigating how to provide ‘dropbox’ type interfaces so the data can be deposited and accessed from a range of devices and in collaborative environments.

Getting to a 'zero cost of use' from a researcher perspective seems to be a key part of strategy to get users to follow policy and not try to go their own way.

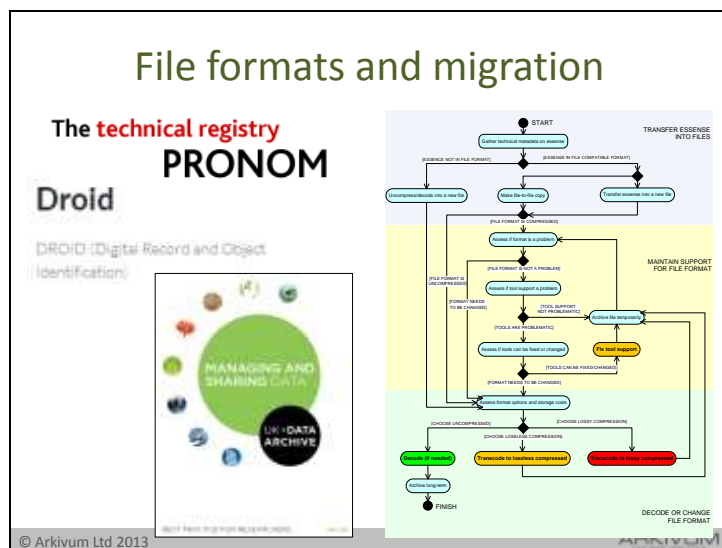
<http://blogs.lshtm.ac.uk/rdmss/rdm-at-lshtm-surveyresults/>



To get a feel of the challenges of metadata, it's worth looking briefly again at the Data Management Rollout project at Oxford [1]. This is a slide that Neil Jefferies from the library presented at a preservation conference called PASIG late last year [3]. It shows the complexity of integrating multiple systems across the University and is built on a series of more specific projects, e.g. DataCite and several others.

Because of the data and metadata integration challenges, there are also real benefits on having a persistent and global namespace for all the research data that is being stored. This means it's much easier to create and maintain catalogues and indexes of where that data is and how to access it – but again more on that later.

[3] https://lib.stanford.edu/files/pasig-oct2012/02-Jefferies_201210PASIG_DAMARO_NJ.pdf



Likewise for file formats, there's a large body of work to draw on, e.g. format identification tools and services such as PRONOM and DROID from the national archive [1], and guidelines on what formats to choose for long-term accessibility.

For example, there's a great guide from the UK data archive [2] on managing research data, which includes format selection. The picture on the right is all about audiovisual data formats and comes from work I did at Southampton in my former life [3] – even for very specific types of data there can still be quite complicated choices to make on what formats to use.

There are lots of ways with dealing with file formats, but the most common strategy is migration to an open, well documented and long-lived format - often when the data is first archived, but sometimes at a later date if the format of the data isn't immediately of concern.

[1] <http://www.nationalarchives.gov.uk/information-management/our-services/dc-file-profiling-tool.htm>

[2] <http://data-archive.ac.uk/media/2894/managingsharing.pdf>

[3] http://eprints.soton.ac.uk/339146/1/339146-_preservationstrategies_R0_v1_01.pdf

Migration: canonical formats

Media type	File formats	Preservation format(s)	Access format(s)	Normalisation tool
Audio	AC3, AIF, MP3, WAV, WMA	WAVE (LPCM)	MP3	FFmpeg
Email	PGC	MBX	MBX	mailkit
Email	Mails*	Original format	MBX	mailkit.py
Office Open XML	DOCX, PPTX, XLSX	Original format	PDF to PPTX	OpenOffice
Plain text	TXT	Original format	Original format	N/A
Portable Document Format	PDF	PDF/A	Original format	Ghostscript
Presentations files	PPT	Original format	PDF	OpenOffice
Raster images	BMP, GIF, JPG, JPEG, PCT, PNG, PSD, TIFF, FLA	Uncompressed TIFF	JPEG	ImageMagick
Raw camera file formats Negative format**	CR2, ARW, ORF, DCR, DCR, DNG, SRF, KDC, MEF, NEF, ORF, PEF, RAF, RAW, SRF	Original format	JPEG	ImageMagick/LRScan
Spreadsheets	XLS	Original format	Original format	N/A
Vector images	AI, EPS, SVG	SVG	PDF	Illustrator
Video	AVI, FLV, MOV, MPEG-1, MPEG-2, MPEG-4, WMV, WMV	FFV1/LPCM or AVC	MPEG-1	FFmpeg

Archivematica 2012

© Arkivum Ltd 2013

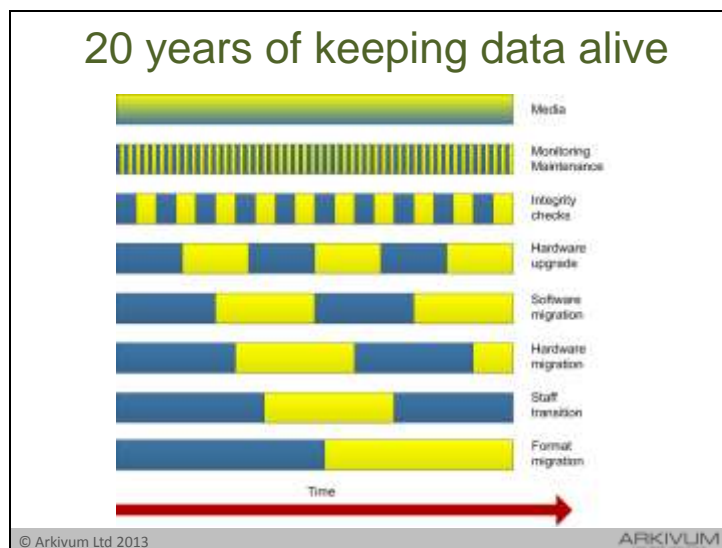
ARKIVUM

This approach of migration of multiple data formats into one or more canonical forms can be seen here in this slide showing what the open-source Archivematica [1] software does in automating the migration process.

The point I want to make here is that format migration is made hugely more difficult if data is held offline, e.g. discs on shelves, or in a range of disparate storage servers where it becomes a management challenge to track the formats and migrate them when at risk.

Migration and format management is greatly simplified through a common storage service with online access to data and a persistent naming scheme for the data being stored.

[1] https://www.archivematica.org/wiki/Main_Page



Talking of migration then brings me to the issue of storage and access and the second half of this webinar.

I've tried to give a broad overview on the challenges and approaches to RDM, and there's more on the specific areas of costs and in-house v.s. outsource in the follow-on webinars, but what I want to focus on now is the specific issues of long-term data retention and access. After all, if you don't have a good solution to safe data storage in place so you know that the data will be there in the future then little else really matters.

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, even long-lived formats such as PDF-A will eventually be obsolete as they are replaced with something better and applications no longer provide backwards compatibility.

The key point is that long-term archiving is an active process and there's always some form of change going on. And when change happens there's always a risk that something goes

wrong, and there's always the need to validate that the change has been effected properly. This all requires time, expertise and money. Digital archiving is very different to paper archiving – it's not a case of 'file and forget' rather it's a case of continual interventions to keep content alive and accessible.

Obsolescence

- Obsolescence happens quickly
- 'Media' might have a long-shelf life, but drives and applications that understand that media typically don't
- Ask Andrew Brown...

© Arkivum Ltd 2013

Obsolescence is the main enemy to long-term data retention and happens on worryingly short timescales.

One tendency is to choose long-lived technologies, e.g. archival grade digital storage media, but this isn't typically the best way forward. The media might last 100 years, but the drives to read that media quickly become out-of-date and no longer supported and available.

Just to illustrate this, a guy called Andrew Brown was in the news late last year. He asked his local hospital for a copy of an echo cardiogram that was performed on him in 2004. The BBC [1] reported that the Worcestershire Acute Hospitals NHS Trust said it would cost £2000 for him to be given a copy of his data. The Register reported that the trust said this "was not a cost-effective use of public money" [2]. The hospital does have his echo cardiogram data, which is stored on Magneto-Optical disk, but the hospital no longer has a drive that can read these disks as they have subsequently installed a new archive system [3]. Their supplier apparently said that they didn't stock the drive anymore as it was no longer in production and they would have to ship a drive in from the United States if the Hospital wanted to read the data. That's technical obsolescence in action – and in just over 6 years. This sort of thing really does happen. It illustrates why you need to think carefully whether 'long-lived' media is the right way forward. It's tempting to think that using specialist storage media that's designed to last for decades or centuries is the answer, including magneto-optical disks and other forms of 'archival grade' storage. But this technology often addresses a niche market, which means it can be harder for the companies who develop and sell it to make a sustainable business. The storage media might last for decades, but the companies who make it might not, or they are forced to move on to develop something new.

[1]<http://www.bbc.co.uk/news/uk-england-hereford-worcester-20235193>

[2]http://www.theregister.co.uk/2012/11/08/nhs_scan_2k/

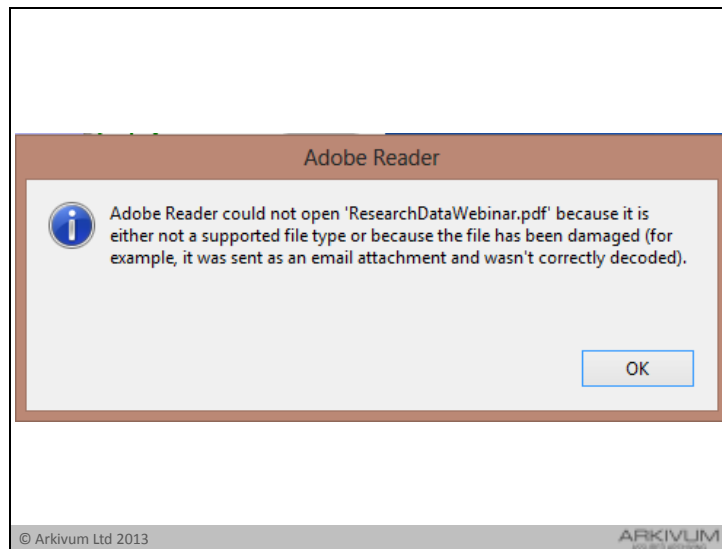
[3]<http://www.whatdotheyknow.com/request/94658/response/237590/attach/2/attachme nt.pdf>



So suppose you go down the route of using commodity IT storage rather than specialised archival media. No IT storage is yet 100% reliable, so there's another set of risks to deal with. Hard drives are an example of supreme engineering, but they do go wrong – as a colleague in the storage industry said 'they are designed to be on the edge of not working'. 1% of hard drives simply fail each year – they won't work at all – and those that do work can suffer data corruption issues – even if used in storage systems (e.g. RAID arrays) that are meant to protect against data loss.

As I mentioned before, if you have data on USB drives on a shelf then worry.

And if you have data on a server, then also worry – but this time about the 10s of thousands of lines of software and firmware code that's in those servers and the bugs that we all know software contains.



And this is what happens when there's a problem.

It's typically 'all or nothing' when it comes to reading a damaged file. A failure often means heroic (and costly) measures to get the data back – if you can at all.

CERN did a test on their storage systems – which had been designed to keep data safe - 33700 files were written to storage and read-back again. (~8.7 TB). 22 were corrupted – and worse still, silently corrupted, i.e. they only knew because they checked each one.

CERN also did a test to see the impact of corruption. They took 10000 compressed files (zip) - 99.8 % wouldn't open if there was just a SINGLE bit error.

<http://indico.cern.ch/getFile.py/access?contribId=3&sessionId=0&resId=1&materialId=pape&confId=13797>

It's not getting any better!

- 1000 times more HDD capacity over last 15 years
- Only 10 times lower Bit Error Rates (BER)
- HDD BER = 10^{-14}
- 1 TB = 10^{13} bits
- 10% chance of an error when reading all of a HDD
- *Within a few years, you are more likely than not to get a read error when copying a HDD*

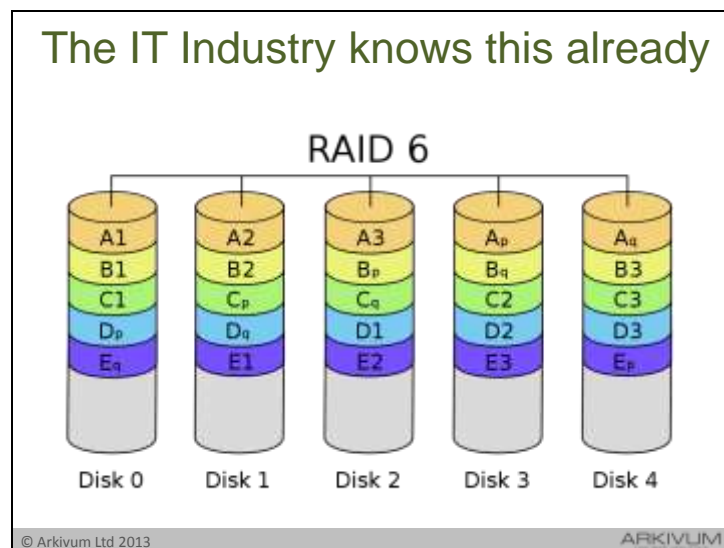
© Arkivum Ltd 2013ARKIVUM

Whilst data capacity increases rapidly for a given cost, error rates aren't keeping pace. Storage costs per TB for Hard drives have fallen by a factor of 1000 over the last 15 years, but error rates have only fallen by a factor of 10.

At the same time, file sizes are going up – for example imaging, time-series, geo-spatial, whole genomes etc.

What this means in practice is that some form of error, even if only small, is becoming increasingly likely in today's ever larger data sets.

http://entertainmentstorage.org/articles/Hard%20Disk%20Drives_%20The%20Good,%20The%20Bad%20and%20The%20Ugly.pdf



This is no surprise to the storage industry, which is why storage vendors continue to develop ever more sophisticated ways to deal with failures and data corruption, for example RAID arrays and more recently self-healing file systems and other fun things.

[1] <http://queue.acm.org/detail.cfm?id=1670144>

Systems bring their own problems

“Disk failures are not always a dominant factor of storage subsystem failures, and a reliability study for storage subsystems cannot only focus on disk failures. Resilient mechanisms should target all failure types”

2008 NetApp study of 1.8M HDD in 155,000 systems

But these systems aren't foolproof – there can be many tens of thousands of lines of code in controllers and file system software – and that means bugs – and those bugs can mean data loss.


Netapp did a big study of over 1.5 million hard drives out in the field in a range of storage servers and found that there failures at all levels and worrying about disk reliability wasn't necessarily the most important place to start.

[1]

http://static.usenix.org/event/fast08/tech/full_papers/bairavasundaram/bairavasundaram_.html/

Risks

- Technical obsolescence, e.g. formats and apps
- Hardware failures, e.g. digital storage systems
- Loss of staff, e.g. skilled archive and IT staff
- Insufficient budget, e.g. storage too expensive
- Accidental loss, e.g. human error
- Selection, e.g. don't retain what you should
- Stakeholders, e.g. retention no longer a priority
- Underestimation of resources or effort
- Fire, flood, meteors, aliens...



© Arkivum Ltd 2013

What it actually all comes down to is the risks are of data loss – and storage media and systems is just one small area.

Many things that can happen that cause loss of content. Some are technical, e.g. storage failures, but some are to do with people, processes, planning and unforeseen events. The key is a holistic approach that applies risk management to drive all areas down to acceptable levels.

https://prestoprimews.ina.fr/public/deliverables/PP_WP3_ID3.2.1_ThreatsMassStorage_R0_v1.00.pdf

Manage the risks

- Do nothing
- Do the wrong thing
- Do it in-house
- Use a service provider



- ISO27001 Information Security Management
- ISO16363 Trusted Digital Repositories

© Arkivum Ltd 2013 ARKIVUM

The biggest risk for preservation is to do nothing - indecision and delayed action.

Then if you do something then there's the question of making the wrong choice.

There's also the issue of whether you do it yourself, which requires the necessary in-house skills and expertise, or whether you outsource to a service provider to take advantage of the knowledge and infrastructure that they may have but you don't.

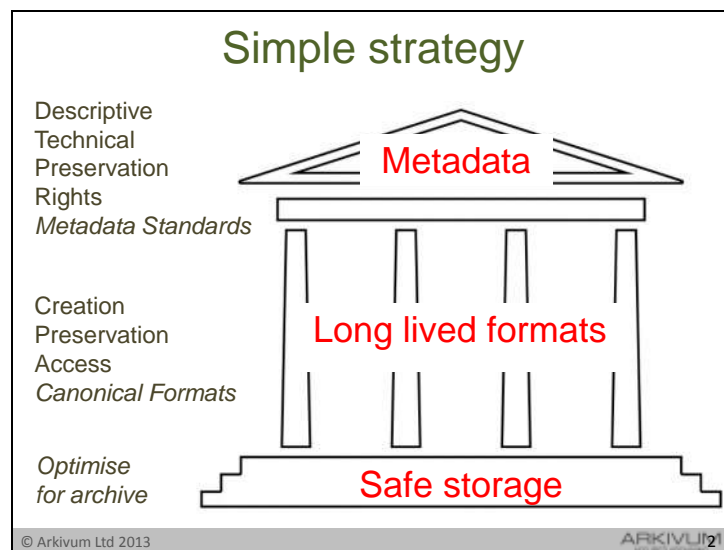
It's all about being informed and managing risk, including managing the risk of using a third-party to provide some or all of the solution.

Beyond regulations, there are two key standards to be aware of: ISO27001 [1] for information security management and more recently ISO 16363 [2] which used to be: TRAC trusted repository audit criteria.

The value is that both standards provide a useful checklist of what to look for in solutions or what to do if you're building your own.

[1] <http://www.27001-online.com/>

[2] <http://public.ccsds.org/publications/archive/652x0m1.pdf>



It's easy to get bogged down in complex solutions that try to crack all the problems at once, e.g. metadata+formats+storage+risk management

A simple approach to get going is to build a 'preservation house'. Start at the bottom and build up.

The key thing is solid foundations – storage and storage management that won't lose your data

Then comes file formats – the approach here is to choose a long-lived format if possible.

Finally, on top of the house is the roof, which means metadata to describes what data is being held and allows it to be found again in the future. Metadata can generally be split into four types:

Descriptive metadata that says what's in the file,
Technical metadata that says what the format is,
Preservation metadata that says what to do, or has been done, to the data to keep it accessible, e.g. migration
Rights metadata that says who owns or can use the data, e.g. IPR

These should all be captured in an open format, e.g. XML, ideally using a standard, e.g. METS and PREMIS.

Three recommendations

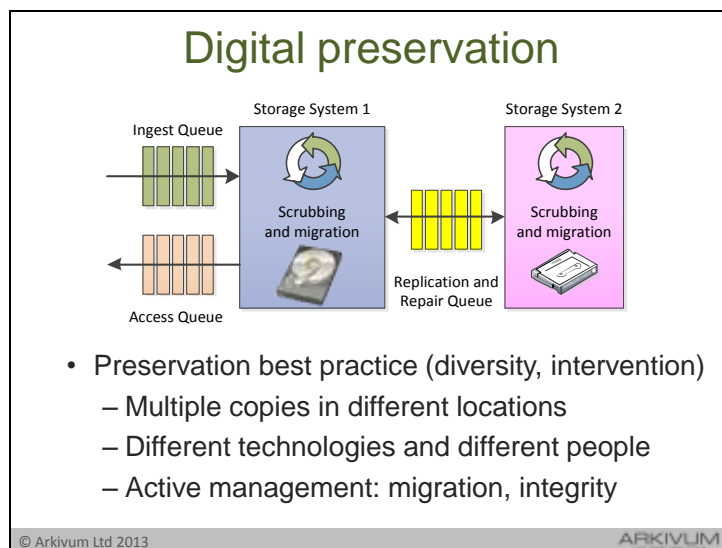
1. Adopt risk management
2. Start somewhere, start now
3. Build a preservation house

© Arkivum Ltd 2013 **ARKIVUM**

So just to summarise, this is what I'd suggest:

- Adopt a risk management approach and recognise the wide range of risks when using IT based systems, processes and people.
- Get going now, don't delay and start with the basics of at least 'keeping the bits safe', and that means using archive storage.
- Build the rest of the preservation strategy on this solid foundation by deciding how to deal with file formats and contextual metadata

By choosing the right approach to storage, especially storage as a service (internally deployed or sourced from a third-party) then the rest can be built on top as the right approach becomes clearer.



So what do you do to get started at the safe storage level – the foundation of the preservation house?

- You make multiple copies of the data and you keep them in different locations.
- You use diverse technologies to spread the risk of failures, which effectively means not all eggs in one basket
- And you actively manage these copies by (a) migrating to new storage or formats to address obsolescence and (b) by regularly checking and repairing any loss of data integrity (which is why having multiple copies is so important – if there is a problem with one of the copies then it can be replaced by replicating one of the other good copies)

Cost v.s. safety v.s. access			
	Data tape on shelves	HDD in servers	Storage as a Service
Storage Cost	Low (media, shelves, climate control)	High (servers, power, cooling, maintenance)	High (fully managed service)
Access Cost	High (people retrieve and load media)	Low (internal network, automated)	High (bandwidth, charges for i/o)
Latent Failures	Low (data tape is reliable)	Med (‘bit rot’)	Low (replication and monitoring)
Access Failures	Medium (people handle tapes)	Low/Medium (depends on system)	Low (automated checks)
© Arkivum Ltd 2013		ARKIVUM	

In the end, you face a trade-off between cost, data safety and ease of access. There is no solution that has low cost, high safety, and instant access. The right approach is to pick a combination that gives the right balance.

<http://eprints.soton.ac.uk/271072/1/AddisIBC2010PaperID298.pdf>

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

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

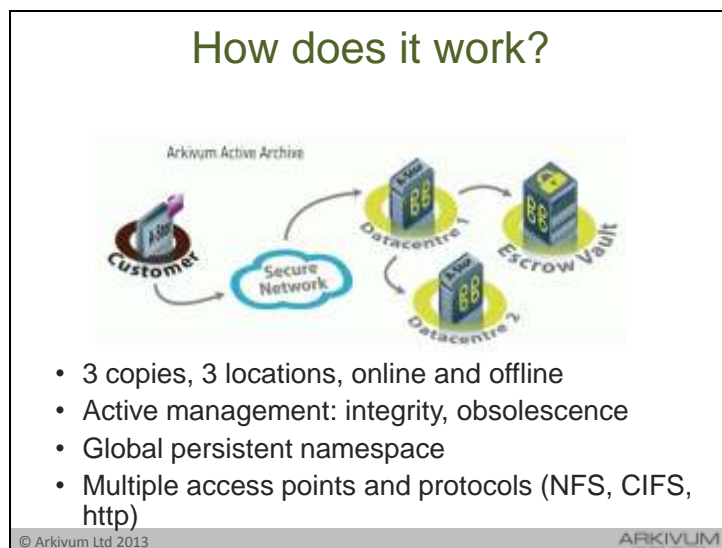
© Arkivum Ltd 2013ARKIVUM

And this is exactly what we do at Arkivum.

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, over the last 10 years on how to retain digital assets for the long-term.

The result is a service that provides safe and secure storage of data that is easy to access whenever it's needed.



Customers use our service through an appliance that sits on their network which provides them with a gateway into our service. The appliance encrypts customer data and generates fixity checksums. Data is then replicated to online data tape libraries in two separate data centres. A third copy is then stored off line with a third-party escrow provider – but more of that in a minute. When we have the three copies in our service then the customer can choose to remove their local copy if they want to.

Three copies of customer data in three locations with two online and one offline allows us to follow data preservation best practice. We use the fixity information 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 and performance remains high.

Not only can we deliver this as a service from our facilities, we can also bundle the whole thing up and install and operate it on the customer site or in the customer's data centre if security of legislative reasons require an 'onsite' solution.

One of the key features is that the solution provides a single, global and persistent name space for data that's stored. It can be ingested and accessed from multiple locations, using standard protocols such as file shares or web. No matter how and where the files are replicated and stored within the solution, the names and paths seen by the users don't change. This greatly simplifies the systems that sit on top that need to deal with metadata, linking publications to data, format migrations and other important aspects of research data management.



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.

100% data integrity guarantee

- All data is returned 'bit perfect'
- No restriction on time
- No restriction on volume
- Included in the SLA
- Worldwide insurance backed: £5M per loss event
- Supported by ISO27001 certification

© Arkivum Ltd 2013ARKIVUM

Coming back to the guarantee of data integrity, the important point to make here is that we have targeted the most stringent of data retention and security requirements from the outset, i.e. we've cracked the toughest of the nuts first. Medical data is a good example and one of our customers is the Oxford Fertility Clinic where we store patient treatment data with a 50 year retention requirement. This means our service will almost certainly be suitable for a very wide range of research data types where requirements are maybe less stringent.

And we have been audited for this, both by ISO27001 but also by our customers, e.g. in the pharmaceutical sector when they do their due diligence on us. It's this independent reassurance that we're doing the right thing that gives us and our customers confidence – and is something that can be harder to confirm for internal university services.

Data escrow

- Copy of data offline at third-party escrow site
- LTFS on LTO tape with open source tools
- Customer has access to escrow copy:
 - If we fail to provide the service
 - If the customer decides to leave



© Arkivum Ltd 2013

And coming back to escrow, because we use data tape, we can create an offline copy of customer data that is lodged with a third-party under a three-way agreement between us, them and our customers.


Use of LTO and LTFS with open source tools means restoring data from escrow is easy and has no lock-in to hardware or software vendors, including us.

Customers can access the escrow copy of their data if we either fail to provide our service or if the customer decides to leave. This gives customer reassurance and an easy exit-strategy if the need it, which is something you don't see with cloud storage providers.

This is really important since whether you have an in-house or outsourced solution, you need to know the exit strategy – how you get out as well as how you get in – research data and the institutions that are responsible for it will outlast almost everything else, especially specific technologies or products – so knowing there is a way to ditch a particular solution if needs be without risking the data is really important – not that I'd want to encourage anyone from ditching us of course!

Pricing

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



© Arkivum Ltd 2013 ARKIVUM

People of course want to know the price of our service.

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.

And we know from the meeting on the 25 April that costs of data retention are recoverable from grants, either through FEC, i.e. overheads for having a data retention facility, or through direct costs, e.g. preparing and storing a specific project dataset.

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.

There'll be a lot more on costs and cost recovery from funding bodies in the third webinar.

Thank you

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

"Arkivum has helped us to create a robust archiving solution that will allow us to focus our budget on the business rather than yet more storage"

"Archived documents can then be seamlessly accessed from within the document management system, in the same way current documents are".

© Arkivum Ltd 2013

ARKIVUM

So this brings us to the end of the webinar. I've tried to highlight some of the many challenges of long-term retention and access to research data. Arkivum doesn't provide a solution to all of these. We do however provide a great place to start and a solid foundation.

With so much to worry about, transferring the responsibility of long-term storage to Arkivum and its experienced team means your resources can be devoted more core issues, be it getting more research done or trying to crack other tough nuts such as metadata or affecting cultural change.