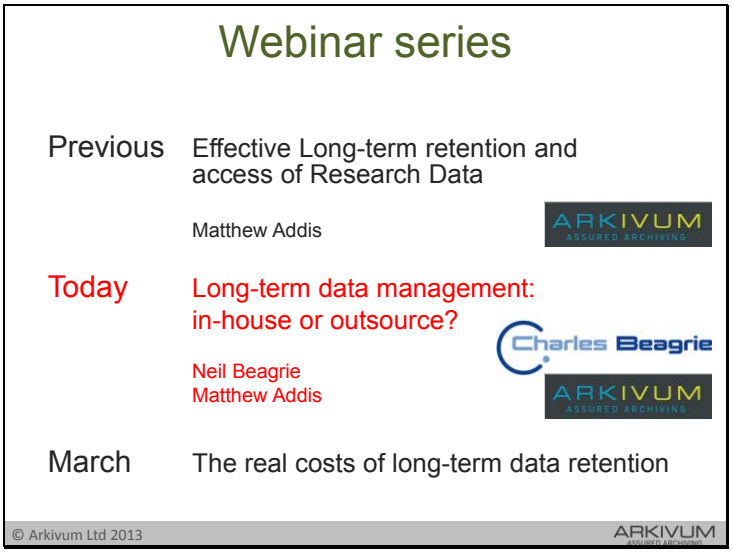
Slide 1

Welcome to today's webinar on long-term research data management, which will focus on the question of in-house or outsourcing.

We have long-standing expert in digital preservation and RDM, Neil Beagrie, to provide a vendor independent viewpoint and give his expertise on the question of shared services and cloud models.

Slide 2



This webinar is the second of three webinars on Research Data Management.

1. The first webinar was an overview of the RDM landscape and issues.

2. Today will be an exploration of the pros and cons of in-house or outsourcing.

3. In the third webinar we'll explore the real through-life costs of data retention and access, including what makes up the TCO and strategies for reducing, sharing and recovering those costs.

This is actually a re-run as we're doing the whole series again because people found them useful first time round. Slides for all the webinars are already online plus recordings of the broadcast. We'll update the slides and recording from today.

I'll now hand over to Neil to get us started.

Slide 3

Overview

Advantages/Disadvantages of Out-sourcing long-term Data Management via 4 aspects:

1. Specialist service
2. Access
3. Deployment
4. Funding and Costs

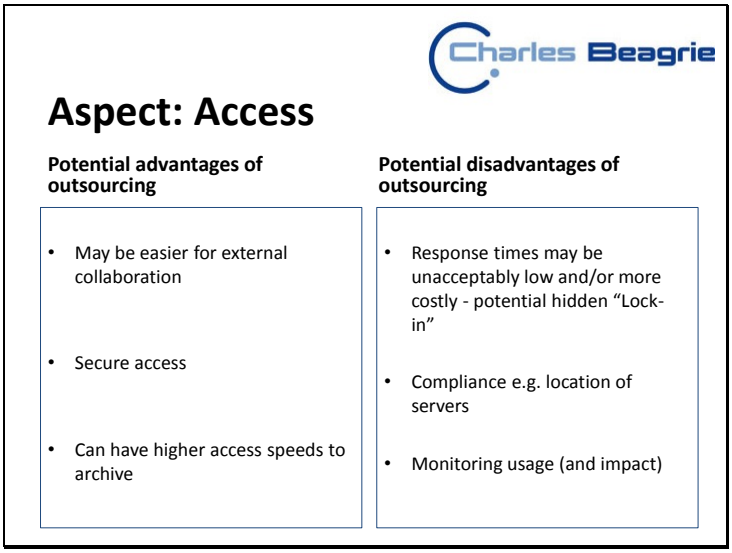
Caveats

1. Pros and Cons are 'potential'
2. Can vary between providers
3. Issues may also apply to long term DM in-house

Slide 4


Charles Beagrie
Aspect: Specialised Service
Potential advantages of outsourcing
• Provides specialist skills, experience and infrastructure which may not be available within the institution
• Allows the institutional IT to focus on other aspects of service provision
• Allows research team to focus on the research
• Data protected from loss
Potential disadvantages of outsourcing
• You need some practical experience and expertise to develop and monitor effective contracts
• Risk of business failure (or success!)
• Very long-term preservation needs knowledge of content and preservation planning

Slide 5

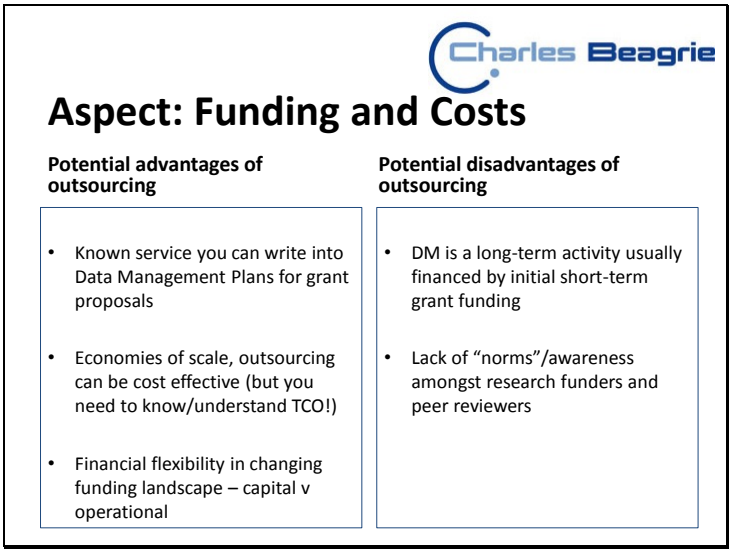

Charles Beagrie
Aspect: Access
Potential advantages of outsourcing
• May be easier for external collaboration
• Secure access
• Can have higher access speeds to archive
Potential disadvantages of outsourcing
• Response times may be unacceptably low and/or more costly - potential hidden "Lock-in"
• Compliance e.g. location of servers
• Monitoring usage (and impact)

Slide 6

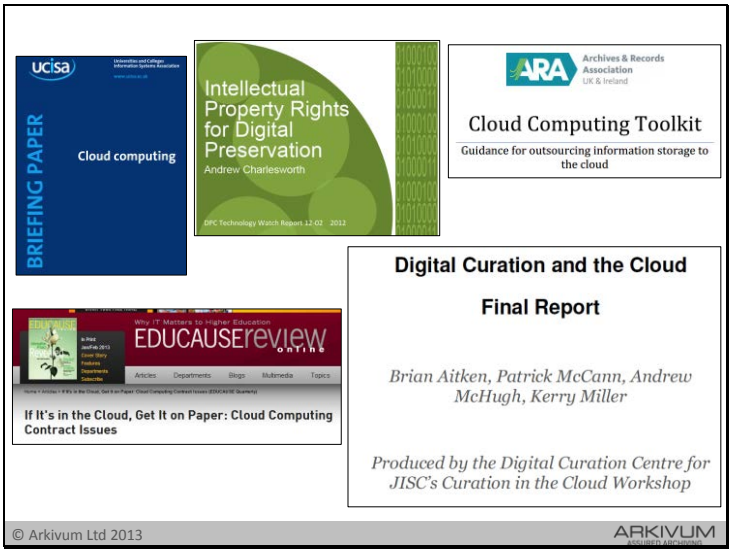
Aspect: Deployment

Potential advantages of outsourcing	Potential disadvantages of outsourcing
• Agility and speed of deployment – service in place when you need it • Flexibility – room to experiment and change things later if you wish	• Connectivity – dependent on internet

Slide 7


Charles Beagrie
Aspect: Funding and Costs
Potential advantages of outsourcing
• Known service you can write into Data Management Plans for grant proposals
• Economies of scale, outsourcing can be cost effective (but you need to know/understand TCO!)
• Financial flexibility in changing funding landscape – capital v operational
Potential disadvantages of outsourcing
• DM is a long-term activity usually financed by initial short-term grant funding
• Lack of "norms"/awareness amongst research funders and peer reviewers

Slide 8



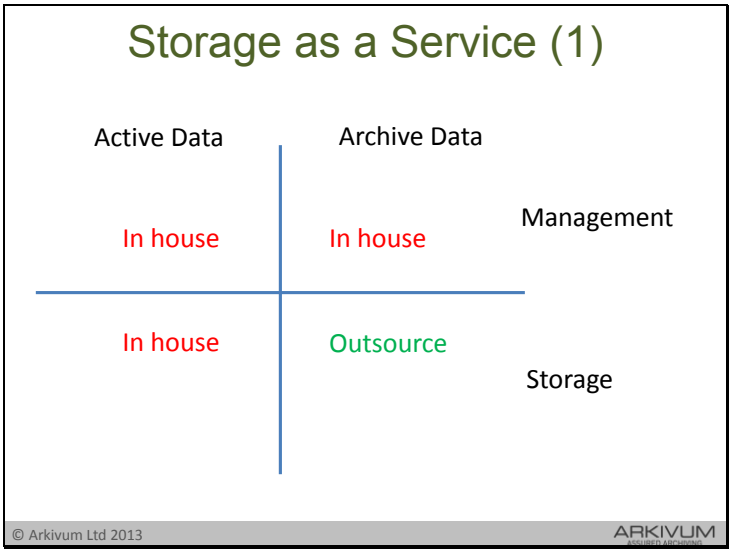
I did think about building on Neil's excellent review of the pros and cons of in-house or outsourcing by going into more detail of the benefits and issues, especially for 'cloud' services which is a hot topic at the moment, but there's already some very useful articles and reports in this area which do a fine job. Here's just some of the examples.

There's a good general overview by UCISA and similar reports from outside of the UK, in particular a really detailed examination by EDUCAUSE who have a series of articles on cloud. And then there's more specific material in the area of archiving, curation and research data, for example a report from the Archives and Records Association and the DCC/JISC, both of which are well worth the read.

Finally, and whilst not specifically about outsourcing, there's a good report on IPR and legal issues in digital preservation which is of course important to consider when considering a third-party service. This is important, when making sure contracts address issues around what actions a service provider is allowed to take that might otherwise be prohibited by copyright, for example format migrations, maintaining or reverse engineering applications, dealing with content where ownership is undetermined or joint between multiple institutions.

http://www.educause.edu/ero/article/if-its-cloud-get-it-paper-cloud-computing-contract-issues
http://www.archives.org.uk/images/documents/Cloud_Computing_Toolkit-2.pdf
http://www.jisc.ac.uk/events/2012/03/curationinthecloud.aspx
http://www.ucisa.ac.uk/publications/cloud.aspx
http://dx.doi.org/10.7207/twr12-02

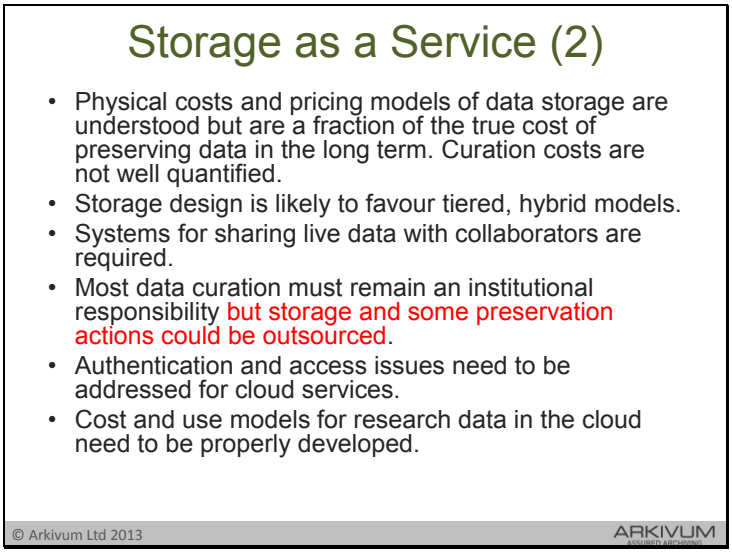
Slide 9



There was a meeting organised by JISC and the DCC on 25 February this year on Storage Requirements for RDM. This was attended by 20 or so Universities who discussed different models and approaches to storage throughout the research data lifecycle.

Arkivum was invited and it was really interesting to see a lot of commonality on what was a candidate for outsourcing. Many of the attendees distinguished between active data, i.e. data that is in the process of being created and analysed, e.g. during a research project, and archive data that needed to be held afterwards. Then there was a distinction between the nuts and bolts of storing data and then higher level functions of managing the data, e.g. curation, integration with a Research Information system (RIS) or Institutional Repository (IR) etc. The management function was viewed as something that should remain in-house, which isn't surprising as there is a lot of university specific expertise and knowledge involved. Likewise for storage for active data, for example to support high performance computing. But long-term storage of data was something that was considered by many as a target for outsourcing.

Slide 10



http://www.dcc.ac.uk/blog/defining-institutional-data-storage-requirements

This is born out in the conclusions of the workshop, which are summarised in a blog post on the DCC website.

There was also a lot of discussion on costs, and what the funding bodies might pay for, but this is something we'll cover more in the next webinar.

Slide 11

http://www.re3data.org/

Using a third-party to store and provide access to research data isn't new of course – there are a large number of repositories out there already – many discipline specific.

The Registries of Research Data Repositories project has just published a catalogue and schema for 338 research data repositories with 171 of these are described by a comprehensive vocabulary. This includes many UK repositories, e.g. ADS and UKDA.

The interesting thing about re3data is the schema used to describe the repositories, e.g. whether there is support for persistent identifiers, what licensing is used for access to the data, and whether the repository has been certified in some way or follow standards.

Slide 12

Many of the repositories in re3data are about publicly accessible data sets, which is of course tied into the current movement towards Open Data as well as Open Access to publications.

Last week there was also a really good set of events hosted by JISC at Oxford on the now and future of data publishing. This included many of the major publishers talking about their support for, or integration with, services for storing and access research data. Open Access was a big theme of the event, which ties into funding body requirements for all outputs of publicly funded research, including research data, to be as widely accessible as possible as they are considered a public good.

There is now a whole ecosystem springing up in this area, e.g. for creating DOIs, storing data sets online and open access publishing.

These are all examples of hosted services from third-parties that are effectively provide outsourcing of some aspects of research data management.

This isn't to say that they replace internal equivalents within a University since they can of course be used in addition rather than instead of in-house solutions.

Slide 13

But the thing for me is that given all the perceived benefits of outsourcing, and existing services, why does uptake seem to be quite low?

Especially for long-term storage of research data, in particular high-volume or high-value primary research data.

What's stopping people realise these benefits? Here I think that there are four main factors at play.

The first is cost. If cost of outsourcing is lower, or if the charging model is a better fit to IT budgets or recovery from grants, then outsourcing can be attractive. Cost saving is a major driver for Universities at the moment. But cost calculations need to consider all costs, which in the case of using cloud services includes bandwidth of getting data to and from the service, which is a common concern, but also less direct costs, for example training people so they know how to negotiate and define a contract and SLA with a provider. But many institutions don't properly account for their full internal costs – they can consider staff, power and space as somehow free – which means they calculate a low in-house cost and compare this to outsourcing and immediately rule out cloud as too expensive when in fact that isn't necessarily true. This is why we want to run a third webinar just on costs.

Then comes the issue of risk. Security and reliability are two commonly cited concerns with cloud services, which includes network connectivity. The risk of outsourcing, esp. for retention and access to valuable assets, will often need to be lower than in-house and often by some considerable distance to even get a look in. I think it's amazing how many risks people take inside their organisation, e.g. using a broom cupboard for a data centre with

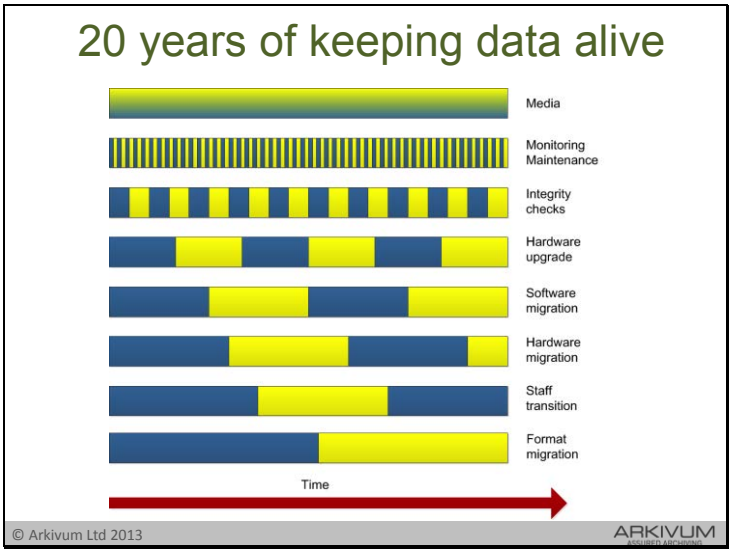
USB drives and no backup, but when it comes to using a service provider the risks have to be driven down to zero – which in turn increases costs and fuels the 'not economic' argument.

Then comes the issue that outsourcing often needs to add value, for example provide functionality or services not available from an in-house solution, such as easier data sharing or better regulatory compliance. It's not enough for the outsourced option to the same as the in-house solution, it needs to allow you to do more.

Finally comes the issue of people. Sure outsourcing provides the opportunity for people to focus their efforts on core business, e.g. doing research, but that's not always how it's perceived. Staff inside an organisation need to feel secure that outsourcing won't jeopardise their jobs – it needs to allow them to do something else that's better paid or that provides career enhancement. The human factor around job security shouldn't be underestimated, esp. in these times of austerity.

Ticking one of the four boxes isn't enough to get over the hurdle of moving from in-house to outsourced models. There can be a lot of inertia when moving to an outsourced model – it will often have to provide more, at the same or lower cost, with less risk, and not putting jobs at risk – unless of course you have no in-house solution at all and no one to build and run one!

Slide 14



In the context of long-term retention and access to research data, I want to start looking at these issues by going back to a slide I used in the previous webinar.

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 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 the start-up costs, and that applies to in-house or outsourced model.

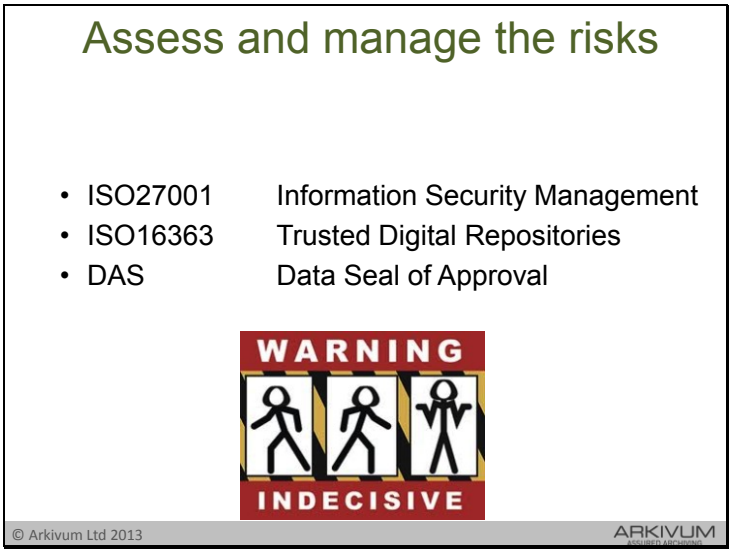
For the issue of risk, it's a case that every time something changes, which is all the time, then there's a risk that something goes wrong, or there's not the money to keep going.

For the issue of value, then one key feature of outsourcing is that it can take a whole world of pain off your hands in having to do all this yourself.

And for the issue of people, it's knowing you have the expertise and staff in-house or not – and whether you can retain this.

With IT costs typically being 2/3rds people, there's a real issue around skills and staff shortage in Universities. We've spoken to Universities who've secured capital budgets to buy kit, but have then spent over a year just trying to find someone to install and run it. Outsourcing can be as much about buying access to people as it is buying access to infrastructure.

Slide 15



And the issue of people comes up again when looking in more detail at risk. There are many risks to research data over decade retention periods, and there are whole frameworks to help assess these risks, e.g. DRAMBORA or TRAC, (and if you aren't familiar with these, then pay a visit to the DCC), but the major risk is that you don't have the right people in-house.

And buying in these people is where the cost goes up – and if you divert other resource, e.g. researcher time or core IT services then something else has to give.

So risk, cost and people are all connected. The FEC cost of the staff needed to keep data safe and make it accessible can be significant.

One approach when considering risks, and the associated costs including people to mitigate them, is to do it in context of what needs to be achieved, e.g. RC requirements or regulations.

We covered RC requirements in the previous webinar, along with some of the other regulations too, e.g. GRP and GCP for lifesciences, or the need to comply with DPA or FOI legislation.

But what's really interesting is that irrespective of whether you're looking at MHRA GCP or RC retention and access policies, it typically all comes down to four factors, and these are typically explicit in the all the guidelines and regulations:

Integrity (can you show the data is still intact), Availability (can you access it when you need it), Confidentiality (can only the right people access it) and Responsibility (who's doing what).

These tie in very well to some of the key standards to be aware of: ISO27001 [1] for information security management and more recently ISO 16363 [2] which used to be called TRAC trusted repository audit criteria. There's also the more lightweight Data Seal of Approval for self-assessment with peer review.

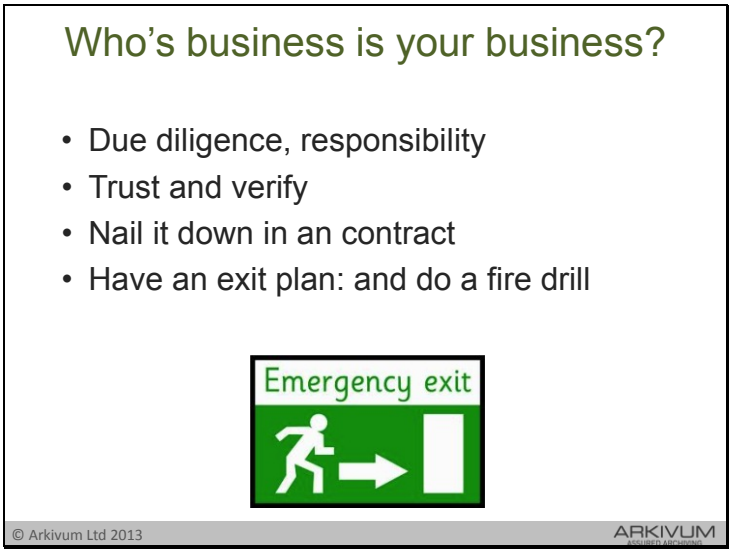
These all provide a useful checklist of what to look for in solutions or what to do if you're building your own.

ISO16363 is interesting, not only because it comes from a set of space agencies that take data retention very seriously, but in that it covers issues such as financial viability and continuity of the business providing a TDR as well as the measures in place to protect the objects it stores.

You won't find many service providers who've got these stamps and not many national repositories either (except UK data archive). This is largely a reflection on how cloud services are very much about IT infrastructure rather than research data management. But the standards provide a very useful assessment and comparison tool.

[1] http://www.27001-online.com/
[2] http://public.ccsds.org/publications/archive/652x0m1.pdf
[3] http://datasealofapproval.org/

Slide 16



One approach to keeping the risks down when using a service provider is to, as Ronald Reagan said in the cold war, trust but verify.

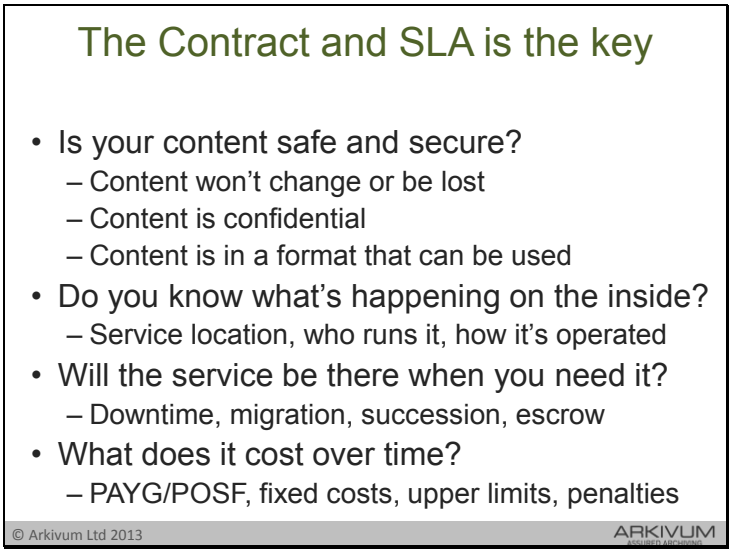
Find a service provider whose business is aligned with and dependent on your business, i.e. they will work in partnership with you. This is important because there are many storage service providers out there, but few have a core business on data retention, so will they really be that concerned with giving you the best service if you are just a marginal part of their business?

Then do due diligence to make sure you can trust them, but get everything nailed down in a contract in case it doesn't work out.

And most importantly have an exit strategy.

Think first about how easy it is to get out before working out how you will get in. And test this exit strategy – do tests to see if you can get your data back or move it to another provider.

Slide 17



So looking in more detail about the contract and SLA with the service provider, then this could be outsourcing, but should really apply equally to a service provided within an institution.

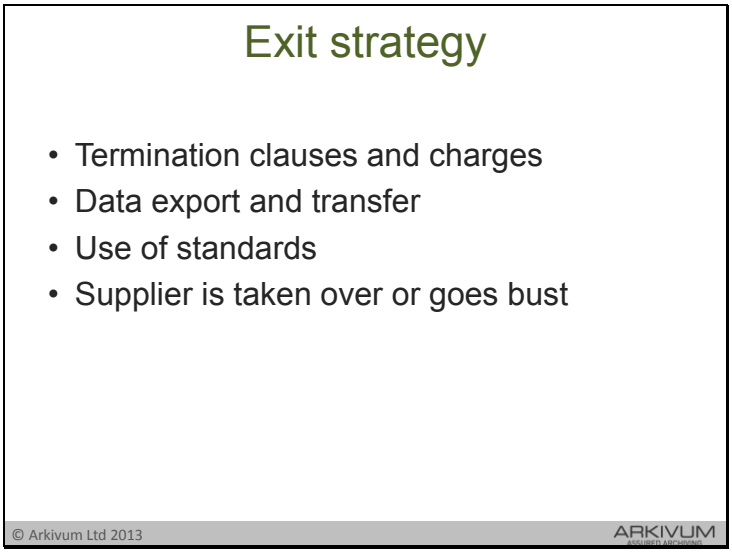
1. First, is content safe and secure in the service? Are there assurances that fixity is maintained, security measures are in place to prevent unauthorised access, and will formats used allow the content to be read and understood in the future. These aren't just technical issues, but include the people who operate the service and the processes and procedures they follow.

2. Second is whether the service is transparent, do you know who operates the service, where the content is stored, what actions and interventions are made, and is this reported? This is about possession, authenticity and provenance.

3. Third comes the question of availability and sustainability, not just in the IT sense of how many hours of the day the service will be running, but what measures are in place to keep it running, including migrations to address obsolescence and succession or escrow if the service provider gets into difficulties or you want to swap providers.

4. Finally is cost, in particular can you determine what the costs will be in the future as well as today, are the costs bounded, what penalties/pay-outs/rebates or ring-fencing happen if either side defaults on their obligations.

Slide 18



Getting an SLA nailed down is great, but perhaps the most important thing is to have an exit strategy. You only expect to use the contract when things go wrong. But getting embroiled in contract breaches, remedies, lawyers etc. is no fun – and gives no guarantees of getting data back or service restored.

So you need a tested exit plan.

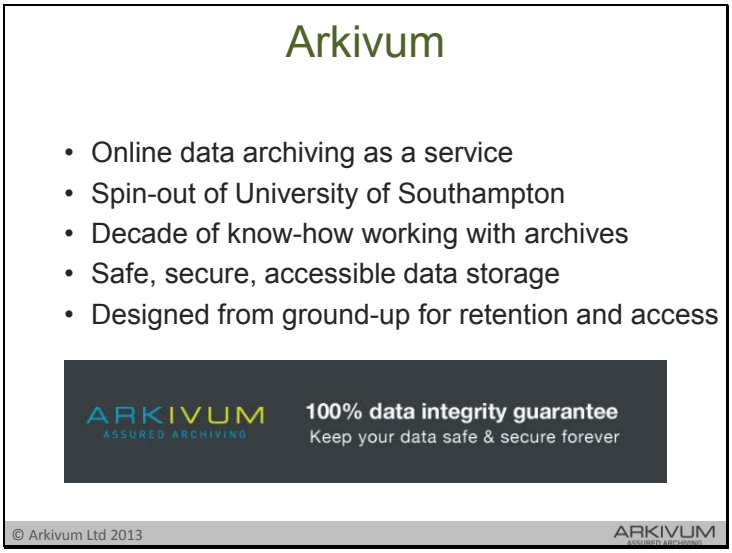
This means worrying about how you can terminate a service or in-house solution and what it costs.

How can you get data out of one service or solution and into another – and this depends hugely on use of formats and standard?

And ultimately what happens if a supplier goes bust, be it a vendor of an in-house solution or a service provider.

Again it's the contract that's key to this – as well as contingency plans and testing the exit strategy really will work.

Slide 19



So now a few slides and the last 5 minutes on how Arkivum addresses these problems.

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.

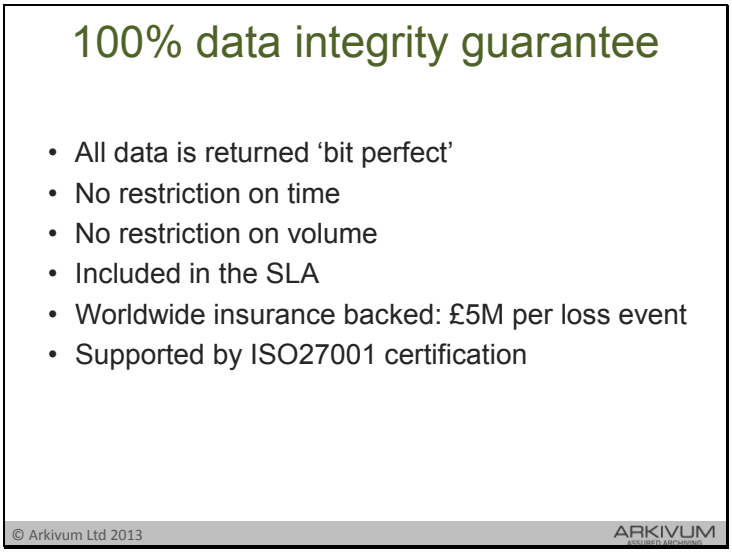
Slide 20

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 integritythrough 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.

Slide 21



Good preservation practice based, trained staff and very carefully controlled processes means we can offer a guarantee of data integrity.

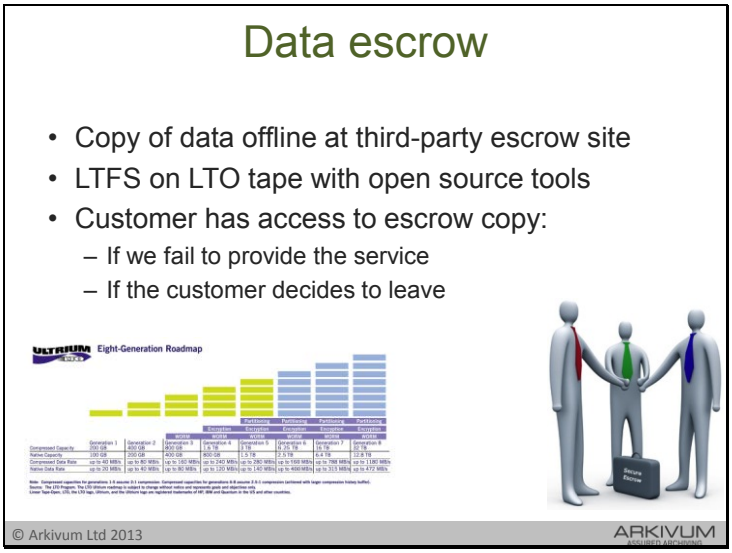
All data returned from our service is always bit-for-bit identical to the data the customer supplied, with no restrictions on time or volume.

The guarantee is backed by insurance, is included in our SLA. Furthermore, Arkivum is certified to ISO27001 where Arkivum has been audited for the integrity, confidentiality and availability of data assets in our possession.

This is where we address the 'value' issue by providing something extra – in this case guaranteed data safety - but also the 'people' issue of having staff that can deliver a regularly audited solution – which in turn means we address the 'cost' issue of not needing this to happen in-house.

Our service is based on risk management, hence ISO27001, which also means we fit well with TDR models, regulation and compliance requirements, and RC policies. So the 'risk' issue is pretty good. Medical data is a good example and one of our customers is the Oxford Fertility Unit where we store patient treatment data with a 50 year retention requirement.

Slide 22



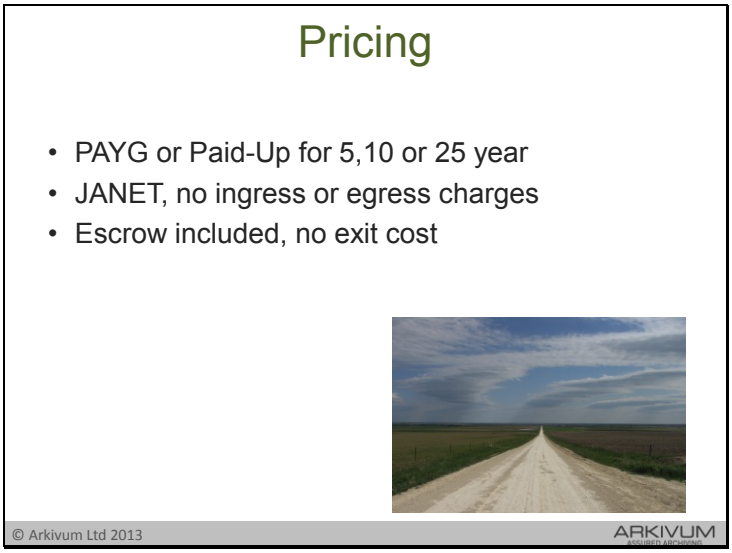
And coming back to escrow, because we use data tape, we can create an offline copy of customer data that is lodged with a third-party under a three-way agreement between us, them and our customers.

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 to ditch us of course!

Slide 23



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.

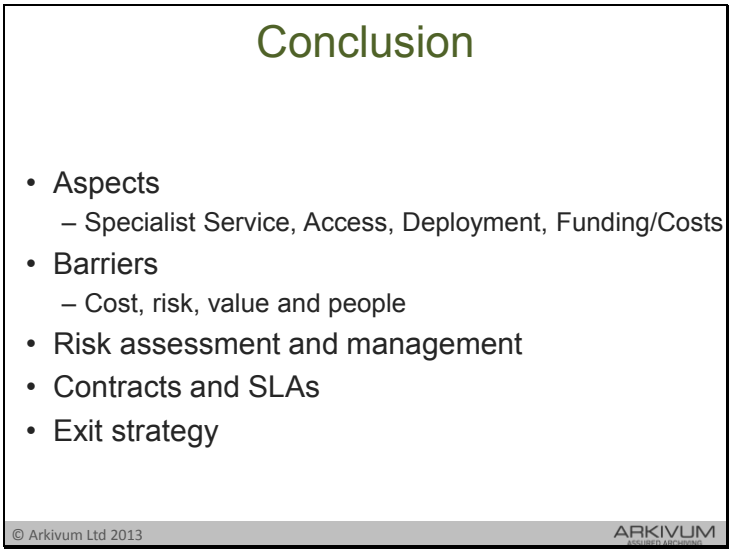
We'll be on JANET by the end of the quarter and we don't charge for ingress or egress of data.

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.

Slide 24



I hope we've shown that the choice to go in-house or outsource comes down to considering all the aspects that Neil talked about, and paying attention to some of the barriers that make it hard to implement change.

The approach I'd advocate is to use risk assessment and management, based on existing standards and models, and to match this to the business requirements of what needs to be achieved, e.g. the RC requirements, regulatory compliance, or the need to provide better access to research data.

When the right way to go is clear, then nail it all down in contracts and SLAs, which includes internal services as well as outsourcing.

Then finally make sure you have a clear exit strategy, succession plan or whatever you want to call it to ensure business continuity.

Of course, we think that Arkivum's solution addresses all of these areas, but even if Arkivum isn't for you, hopefully the things we've talked about will help whatever solution you go for.

Slide 25

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

www.beagrie.com
neil@beagrie.com

"Great Results, Personable, Expert. Neil worked with me on the topic of managing research data. He did a wonderful job of enabling us to understand the key issues, and did a really good review of our case studies. I was hugely impressed with his insights and analysis.“

ARKIVUM
ASSURED ARCHIVING