



Welcome to this webinar on integrating archive data storage with institutional repositories.

I'm Matthew Addis, CTO of Arkivum, and I was previously at the University of Southampton for 16 years working on collaborative R&D projects with Industry on a range of topics including digital preservation and data archiving, as well as eScience, Laboratory Information Systems and other fun topics.

Today I'm joined by Wendy White, who is Head of Scholarly Communication at the University of Southampton where she leads the co-ordination of cross-service research data management support. Wendy has been involved with a range of JISC projects relating to initiatives in open access, repositories, research data and digitisation.

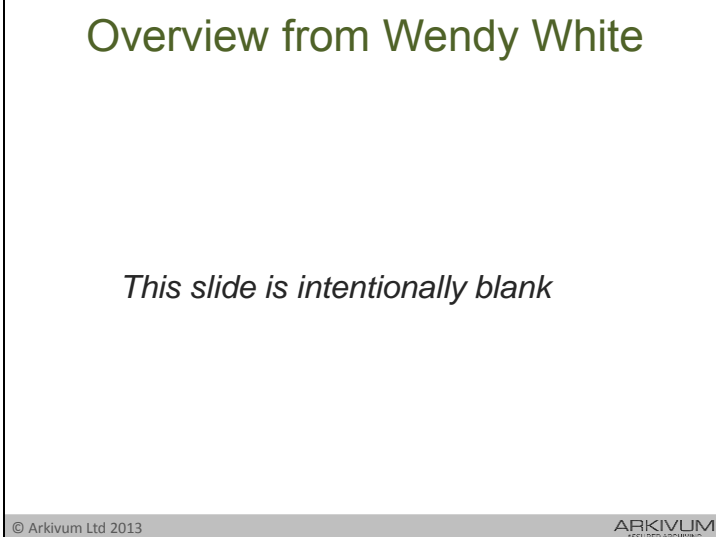
Wendy will be starting the webinar with a high level overview and a focus on the Institutional Repository perspective. I'll then follow by looking in more detail on how archiving and IR fit together with a focus on the archiving perspective.

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
Webinar 4	Integrating archive data storage with Institutional repositories
© Arkivum Ltd 2013 ARKIVUM	

This webinar is 4th in a series that Arkivum has run over the last few months. The previous webinars 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.

1. 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 we explored the costs of data retention and access.

Slide 3



Overview from Wendy White

This slide is intentionally blank

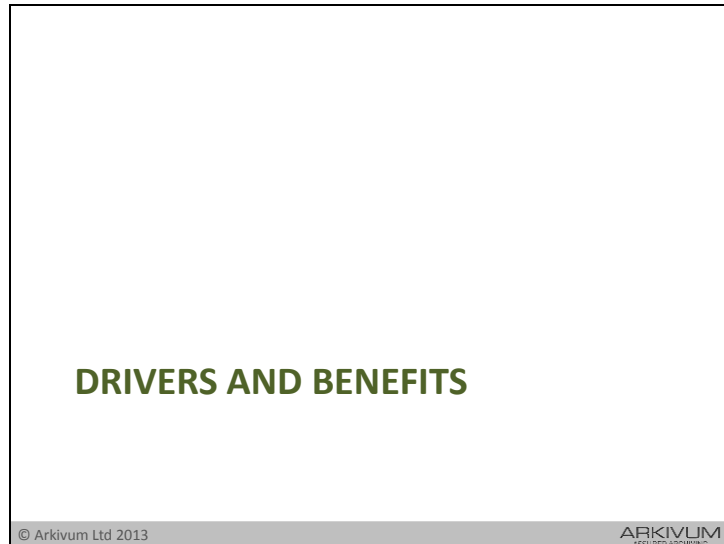
© Arkivum Ltd 2013

ARKIVUM

No speaker's notes are available currently for Wendy's overview, but it can be listened to by going to the Arkivum website and following the link for the webinar.

Contents

- Benefits of archive storage
- Drivers for keeping research data
- Data sources, sinks and lifecycles
- Integration of data archive and Institutional Repository



In this section we look at some of the current motivations for keeping research data safe and access and the benefits to using a combination of an institutional repository and archive to do this.



Research data needs storage, but the type of storage depends on at where the data is in its lifecycle.

The type of storage required when the data is being created, processed and analysed will typically need to give fast access to the data and allow the data to change, e.g. editing, transformations and updates. In some cases there is the need for the storage to support high performance computing applications.

This can be very different to the type of storage that is best suited for long-term storage of data. Here data is no longer going to change and access will be a lot less frequent. The characteristics of high data safety, immutability and manageable long-term costs come into play.

Archive storage

- Low cost
 - A way to cope with data growth
 - Free up expensive resources
 - Reduce the 'cost of loss'
- High safety
 - Data: immutable, replicated, managed
 - Requirements: reuse, compliance
- Occasional access
 - Frequency: each year, not each second
 - Speed: minutes, not in milliseconds

© Arkivum Ltd 2013

ARKIVUM


Archive storage is ideal for research data that needs to be kept for compliance and reuse, esp. primary data which remains static and isn't necessarily used frequently.

Archival storage is optimised for data safety for the long-term and at a low-cost, but with the compromise that access to the data isn't always instant. But that's typically OK for archive data since by definition you typically don't expect to access the data that frequently - and when you do need to access it, waiting a few minutes is fine – you don't need the millisecond speed of access that you'd want when doing lots of data processing.

Because archival storage is cost effective, it provides a way to deal with the ever growing volumes of research data that needs to be retained by Universities.

Tangible 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 MANAGEMENT

In previous webinars we've discussed the benefits and costs of being able to retain and access research data, so I won't dwell on them here.

Integration of archive storage as part of effective RDM has the ability to enable many of these.


Open Science, Open Data

Open scientific research data should be easily discoverable, accessible, assessable, intelligible, useable, and wherever possible interoperable to specific quality standards


G8 UK
UNITED KINGDOM 2013

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

© Arkivum Ltd 2013



RESEARCH COUNCILS UK



Science as an open enterprise
June 2012
ROYAL SOCIETY

ARKIVUM
DIGITAL PRESERVATION

What I do want to do is look at briefly again at some of the drivers for keeping research data. Why you need to keep research data informs the type of approach that will work best. One of the drivers that is getting a lot of press at the moment is the push for open science and open data. For example, the recent G8 statement by the science ministers about the benefits and requirements for science to be open unless there is good to reason not to. This is on the back of influential reports such as from the Royal Society last year. The drive for retention and access to research data goes back a lot further than that of course, for example recommendations from the RCUK back in 2009. 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 – it's all about the public good of making research data available that's been funded from public money. 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. The important point here is that RDM extends to long-term retention and that should be done with the same level of quality and diligence afforded to the research that created the data in the first place.

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

<http://www.rcuk.ac.uk/documents/reviews/grc/RCUKPolicyandGuidelinesonGovernanceofGoodResearchPracticeFebruary2013.pdf>

https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/206801/G8_Science_Meeting_Statement_12_June_2013.pdf

<https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex>

http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf



The drive towards publicly accessible data from research is all part of the current movement towards Open Data as well as Open Access to publications.

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

This slide shows some examples of hosted services from third-parties – and all are candidates for some form of interaction and integration when it comes to IR and archiving.

<http://figshare.com/>

<http://orcid.org/>


<http://datacite.org>

<http://datadryad.org/>

<http://www.plosone.org/>

<http://databl.wordpress.com/2013/03/27/workshop-report-making-citation-work-practical-issues-for-institutions/#more-600>

Commercial/Confidential Research



Stratified Medicine
Innovation Platform

Pioneering research to reduce aircraft noise.
Rolls Royce University
Technology Centre in
Gas Turbine Noise

© Arkivum Ltd 2013

ARKIVUM
DIGITAL ARCHIVE

But whilst there is a lot of press and activity around open science and open data, and rightly so, there is a lot of data that can't be made accessible in this way.

This includes research that's done in collaboration with industry or has been directly sponsored by industry. Here commercial confidentiality and exploitation potential means it won't be shared - and may never be shared. Just a few examples are shown in this slide.

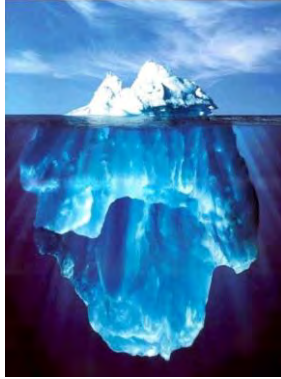
<http://www.cam.ac.uk/research/news/major-investment-will-lead-to-new-materials-for-the-energy-industry>

https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/32383/12-610-wilson-review-business-university-collaboration.pdf

http://issuu.com/university_of_southampton/docs/noiseutc

<http://www.southampton.ac.uk/aerospace/research/aerospace/civilaviationandaeronautics/appliedsolutions/noiseutc.page?>

Bottom of the iceberg



Open Access, e.g. CC0

Legacy data
Commercial research
Personal/private data
Third-party data
Under development
No public repository

© Arkivum Ltd 2013 ARKIVUM

Indeed, there's a lot of data that can't be made Openly Accessible, especially in the strictest definition of the term – i.e. where release to the public domain and no barriers to access exist over the internet.

There may be 591 online repositories that you might be able to put your data in according to Databib.org, but this is only be the tip of the iceberg of research data – much of which still resides inside institutions.

It's the bottom of the Iceberg that's the challenge for an institution. This is the data that has to be kept, for compliance or because it has value - but there's nowhere else to put it. This sort of data can include:

- Legacy data that's not cost effective to prepare for release under open access.
- Commercially sensitive data
- Data that comes under DPA or has ethics issues preventing release
- Data that comes from or has been developed in partnership with a third-party meaning that copyright and other IPR issues prevent release
- Data that is still part of an active research project which may not be ready for publication
- And data that can be released, but there isn't a repository that will willingly accept it, e.g. because of its size or type.

Slide 13

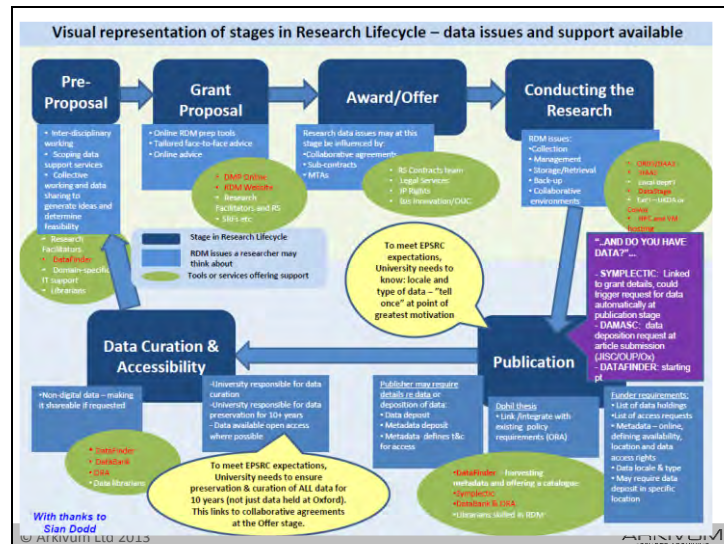


It's this sort of data that we're going to focus on most in this webinar.

So where does this data come from, where might it end up, and what lifecycles does it go through?

Looking at a lifecycle approach helps identify the systems, services and people that an IR and archive might need to interact with.

Slide 14

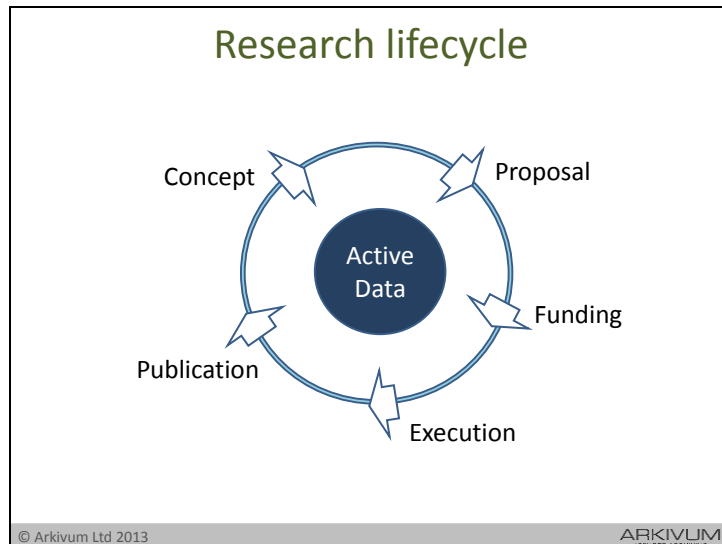


This is a high level view of the research data lifecycle from the University of Oxford.

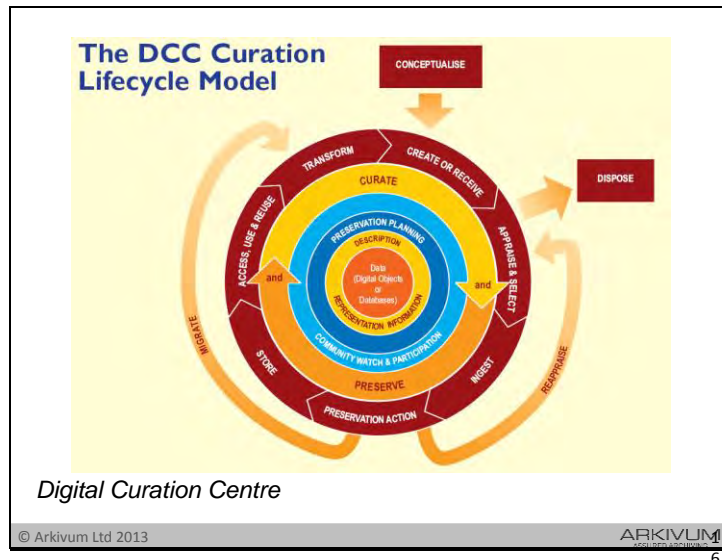
This covers all the stages from initial research idea through to doing the research, publishing the results and then curating and providing access to the data that results.

The lifecycle conflates the ‘research’ part and the ‘curation/preservation/access’ part. An alternative is to treat these as having separate, but interacting lifecycles.

<http://damaro.oucs.ox.ac.uk/>

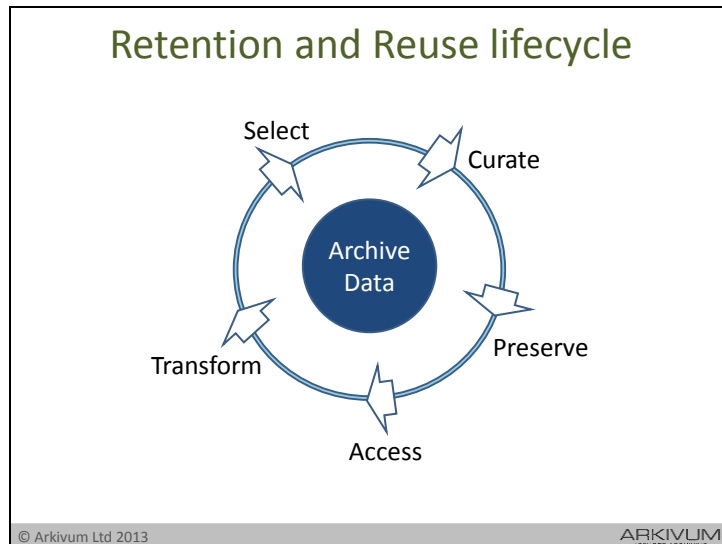


So this is a high level view of the research lifecycle from the concept of a research project through to getting the funding, executing the research and then publishing results. The key thing from a data point of view is that this lifecycle surrounds data that is active, i.e. it is being created, processed, transformed, analysed etc.



Then comes the lifecycle of data that is output from this research process. There are already well developed lifecycle models for curation and preservation, for example this one from the Digital Curation Centre.

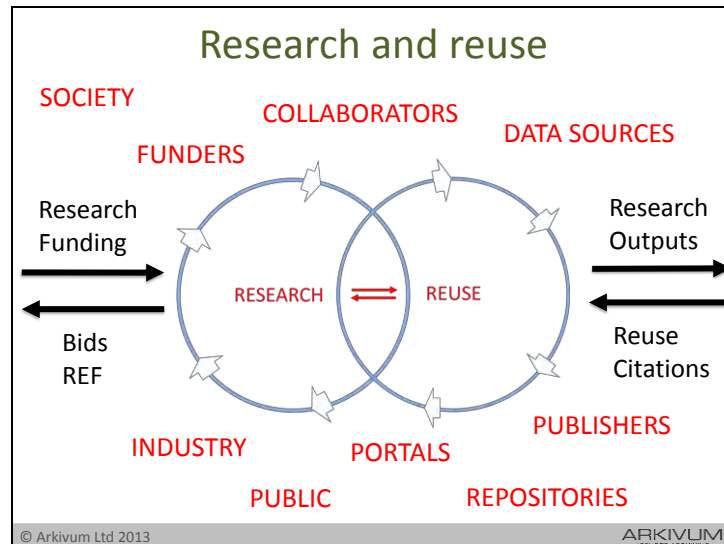
<http://www.dcc.ac.uk/resources>



The curation, preservation and access lifecycle can be summarised as the need to select and curate research data that is created from research projects, undertake preservation activities so it remains understandable in the future, provide access to it for users, and potentially do transformations so it remains usable by the user community.

In this case, the process centres on archive data, i.e. data that doesn't change other than perhaps through preservation actions such as format migrations to ensure that it can still be used.

The primary objective of the lifecycle around archive data is to ensure that it can be reused.

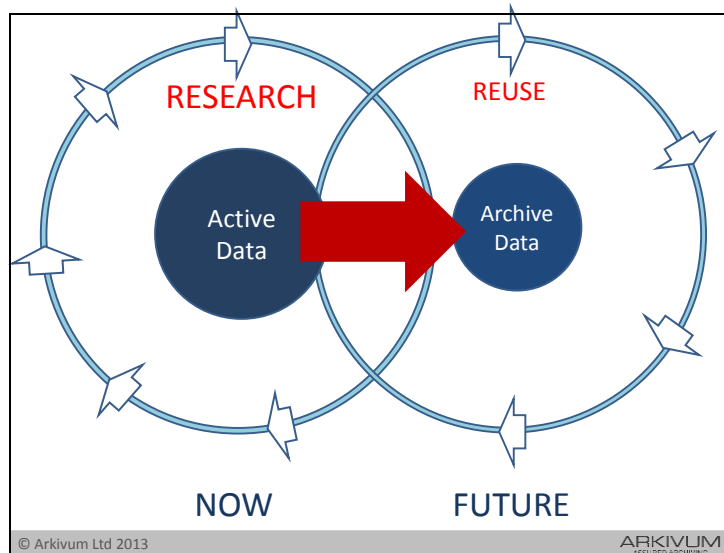


The lifecycle concerned with active data that is created and consumed as part of research intersects and interacts with the lifecycle concerned with archive data that is kept so it can be reused in the future.

Both lifecycles interact with the wider environment, which includes funders, research collaborators, external sources of data, societal drivers such as open science, publishers and portals that provide access to research outputs, external repositories that can be used to deposit data, and even public use of data.

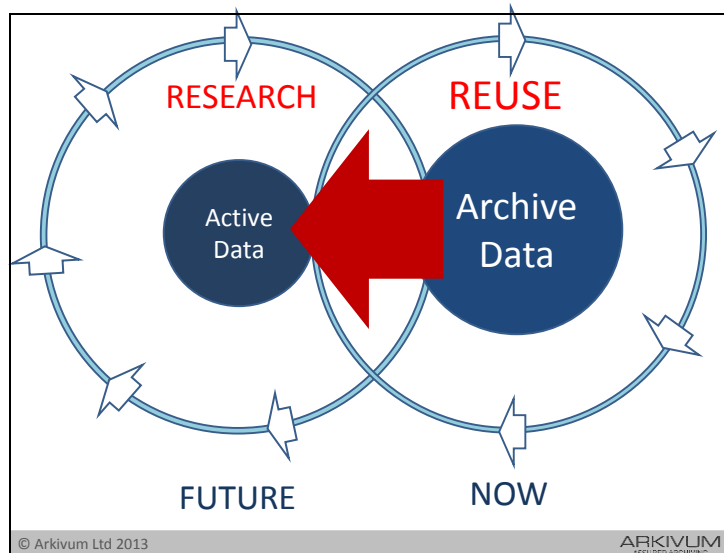
Research money from grant funding pays for research that produces research products. These are made accessible, and get used. Metrics on use then feed back into things like impact analysis of the research and through that to REF submissions for QR funding.

Because there are a lot of stakeholders in this ecosystem, there's many sources and sinks for research data and publications – which in turn means interactions with the Institutional Repository or Archive.

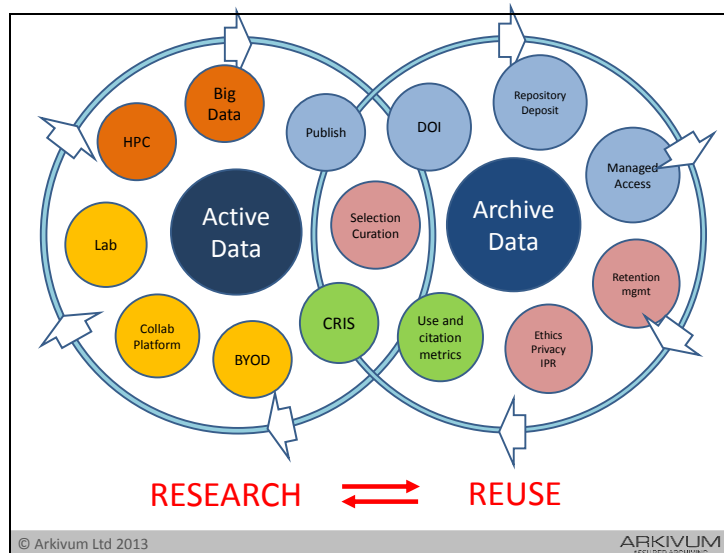


Understandably, the focus is often the research cycle since this provides immediate results and is of course where most of the funding goes. We saw in previous webinars that the cost of the initial research is typically far larger than the cost of curation and subsequent retention.

This creates a continual flow of research data that feeds into the archive.



The archive then grows and accumulates research outputs. All this archive data has value as the basis of new research. This means the flow goes the other way too. As the archive grows then the potential for new research based on previous data will increase. This means the ability to retrieve and reuse archive data becomes increasingly important.



If we look in more detail at the systems involved in each of the lifecycles then we start to get a more detailed view of what interactions are likely with the institutional repository and the archive.

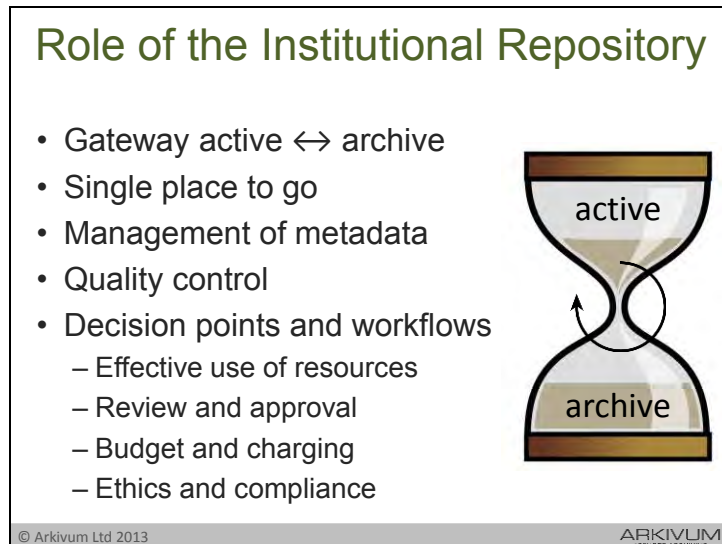
If we take the case that the institutional repository is the place where research data will be initially deposited for safe keeping and subsequent access, then it is the data sources in the research lifecycle that need integration. Integration could be in terms of systems integration for automated data transfer, or integration could be in terms of manual workflows that need to be followed by researchers or PIs. Integration could include:

- Laboratory systems, e.g. LIMS that capture data from equipment used in scientific research
- Collaboration platforms, e.g. document exchange and management where research is being done by several institutions working together.
- Data generated or collected using devices that researchers may themselves supply, e.g. tablets and smartphones used in the field. This is all part of the move towards Bring Your Own Device (BYOD).
- HPC systems for crunching data, running models and simulations
- Big Data analytics and other analysis that might be done on large data sets.

There's a lot of sources of data, many of which have discipline specific platforms and tools associated with them, which makes integration quite a challenge if the desire is to make data acquisition into an Institutional Repository an automated and seamless process.

On the archive side, integration is of course needed with the Institutional Repository, and we'll look at that next, but also comes through interactions with the various services for providing access to data, managing that access, handling issues with providing access, and establishing links and references to the data, e.g. DOIs.

At the intersection of archive and active data comes the systems and services for publishing data and recording its use, with statistics then feeding back into CRIS for example so they can be used as part of research tracking.



The role of the Institutional Repository can be viewed as providing a gateway between the active and the archive data.

It provides a single and centralised place to go to deposit, find and access research data within the institution.

In a sense, the IR is at the neck of an hour glass. The hour glass is wide at the top representing the wide range of sources of research data, i.e. as generated in the research lifecycle. The hourglass is wide at the bottom representing all the systems and services for data access, sharing, preservation and curation, i.e. the archive data. And you can flip the hour glass over to move data between research and archive or back the other way.

The IR is like a managed gateway in-between research and archive that brings some control and order – in particular it allows metadata to be collected and managed as well as the research data. It allows quality control and research governance to be applied. It allows decision making to be integrated at the institution level on how to make best use of resources and budgets.

Recognising the role of the IR and the mediation it may do between producers and consumers of data in the archive informs the interactions that need to be supported between the IR and the archive.



It may seem that we've spent a long time looking at the wider environment surrounding the IR and archive functions, but this is important when defining the interactions between the Institutional Repository and the archive – which is what we look at next.

Interactions: in and out

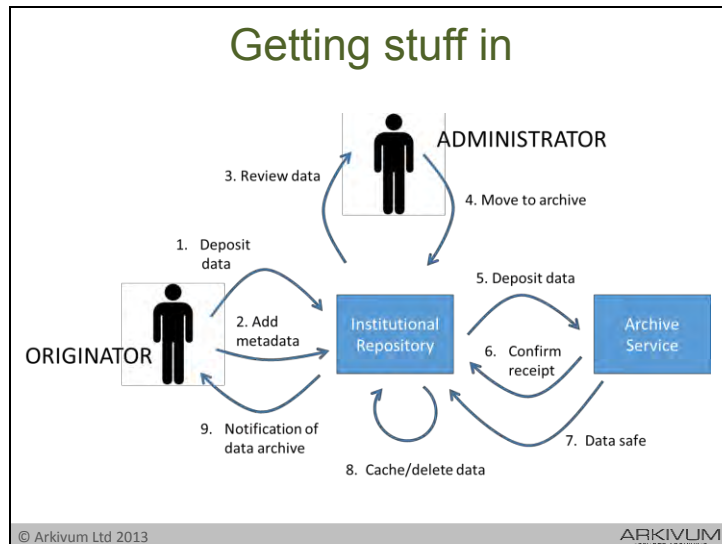
- Getting stuff in
 - Deposit
 - Chain of custody
- Getting stuff out
 - Restore
 - Scheduled and managed use of resources

© Arkivum Ltd 2013

ARKIVUM

The most obvious interactions needed are to get data into the archive and to get it back out again.

But there is more to this than just copying data around between storage systems.



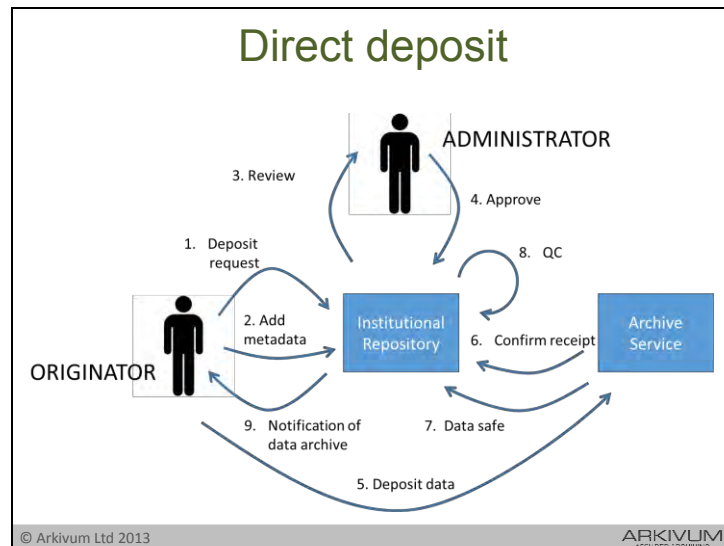
This is an example workflow that might be used.

1. The creator or owner of the data deposits it in the IR, for example this might be done by the PI
2. Metadata is added to the data to provide context. The metadata could include description, authors, restrictions/licenses, embargo/dates and links to a CRIS (funding, projects, locations, collaborators etc.). The deposit of data might be done for example through the recollect plugin to ePrints which supports a workflow for uploading files and metadata, validation and then confirming deposit.
3. When deposit is complete, an administrator is notified.
4. The administrator then decides whether the data should be added to the archive or just held locally in the IR (or both).
5. The IR starts the deposit process into the archive
6. The archive confirms initial receipt of the data
7. The archive confirms that the data is now safe, e.g. after it has been integrity checked and replicated to multiple locations
8. Now that the data has been confirmed as correctly copied to the archive and is safe, then the local copy in the IR can then be deleted if required– or it can be kept in order to provide fast access.
9. It may be that the originator of the data is informed that their data is safe in the archive, although it might also be the case that the decision to archive is done by the institution and the original owner doesn't get informed as the data is not their responsibility at this point.

<http://wiki.eprints.org/w/ReCollect>

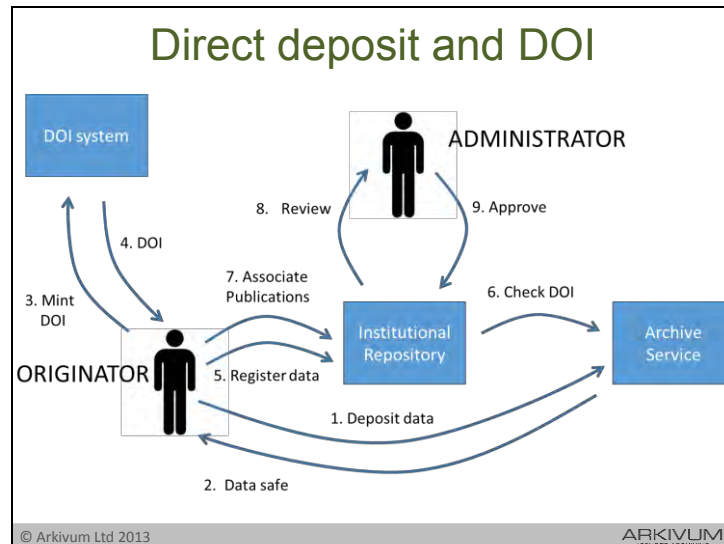
http://www.data-archive.ac.uk/media/395364/rde_march2013_repositoryoutputs.pdf

<http://eprints.soton.ac.uk/352813/3/eprints-sharepoint-report-final10.pdf>



There are several variations on this theme. For example, if the data set to be deposited is very large then it may not make sense to transfer it to the IR only for it to be moved straight to archive.

Terabyte data sets can take days to copy between systems and take up a lot of storage space. In this case, it could be more efficient for the data to be sent straight to the archive system. But it may be that the deposit of the data straight to archive will need approval first. Therefore, in this example workflow, the originator of the data makes a request to deposit data which is then reviewed and approved before they are granted access to the archive in order to do the physical transfer. Since the data hasn't been through the IR, there may also need to be a QC step that checks that the necessary data and metadata are all present before the deposit is confirmed as completed.



The model of direct deposit into the archive could also apply to data that hasn't yet got publications associated with it, or isn't ready for release – for example it is primary data created in the middle of a research project. In this case the data might be directly deposited to the archive, have a DOI created for it so it is uniquely identified and can be referenced, but only at a later date is a record is created in the IR. This workflow is shown in this diagram. Variations on this theme are of course possible.

Archive: chain of custody

- Send data to archive
- Generate checksum for data in archive
- Compare with checksum in IR metadata
- Update metadata for successful transfer
- Apply archive safety mechanisms
 - Replication, WORM, access control
- Update metadata to complete audit trail
- Remove local copy (if desired)

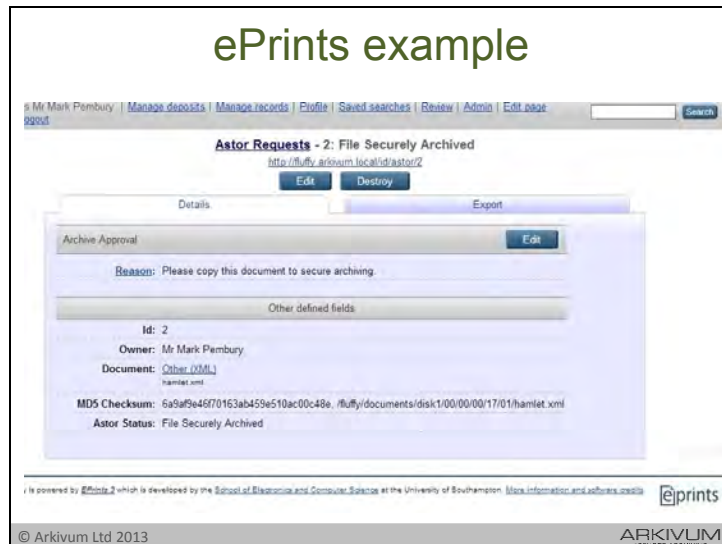
© Arkivum Ltd 2013

ARKIVUM

But irrespective of the model, what you do need is a chain of custody with the archive.

This applies to remote repositories as much as local archive solutions or third-party services.

The question is 'how do you know that data has made it to the archive correctly and is now safely stored?'. The answer is to have a chain of custody. Typically this will include use of checksums or similar measures as the basis of confirming that data hasn't changed during deposit, the application of safety and security mechanisms to ensure that the data is properly protected, and then confirmation that the data is now safe. The 'local' copy of the data should not be removed until this chain of custody process is completed. The steps in the process, especially checksums, then become part of the audit trail that should be maintained to later prove/check that the correct process was followed.



This screenshot shows the results of this chain of custody process when using ePrints to deposit data into the Arkivum service.

A request is made and then approved to copy a document to the secure archive. When the copy is complete then the ePrints metadata for the document is updated to show that the data is now safe. This metadata includes the checksum created by the archive

Getting stuff out

- Open access: no license, no barriers
- Unrestricted use, but request for access
- Unrestricted use, but charge for access
- Restricted use: request/approve access
- Embargoed: no access for set period
- Locked down: no access

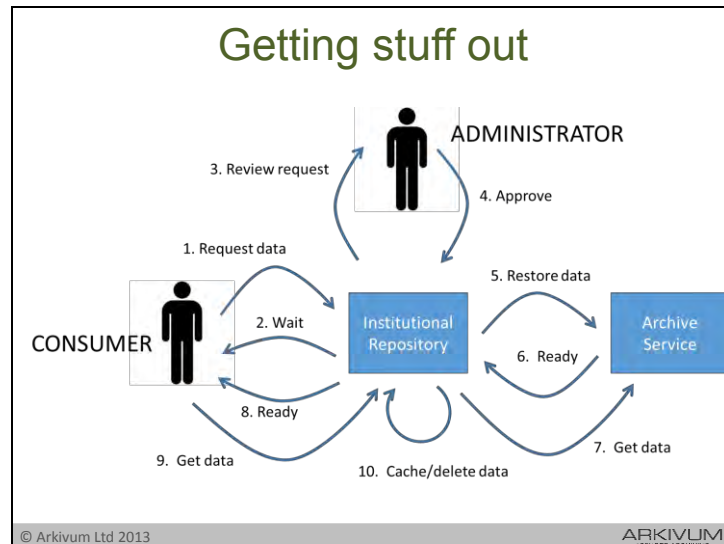
© Arkivum Ltd 2013

ARKIVUM

When it comes to getting data back from the archive, e.g. because of an access request to use the data, or maybe as part of an audit, then there are multiple levels of access that may need to be supported.

These range from open access where there are no request forms, license agreements or other restrictions to access through to data that has to be completely locked down, e.g. because of confidentiality. In the middle comes cases where usage is restricted but possible and a request has to be made to access the data.

Most of the scenarios require a request/approve/deny process. There may also need to be a request/restore/deliver process for the data itself.



Here is an example workflow.

- The consumer of the data makes a request for access
- They are told that they will need to wait for their request to be approved and the data prepared
- The administrator of the IR is notified of a access request
- The administrator reviews and then approves/rejects the request
- Approval triggers restore of the data from the archive.
- For large data sets this could take hours or days so the IR is notified when the data is ready.
- The IR retrieves the data from the archive.
- The consumer is told that the data is ready
- The consumer retrieves the data
- The copy restored to the IR can be deleted or held for longer if it likely that another request for the same data will come in

The review/approve step could involve other stakeholders, e.g. ethical or legal teams if there are privacy or contractual issues to address.

Review/approve

- Large data sets
 - Time, e.g. restore from archive
 - Money, e.g. storage and bandwidth
 - Effort, e.g. packaging and delivery
- Private or confidential data
 - Who, why, when, how
 - Authentication
 - Controlled access
 - Licensing
 - Audit trail
- Interaction with consumer ≠ open access

© Arkivum Ltd 2013

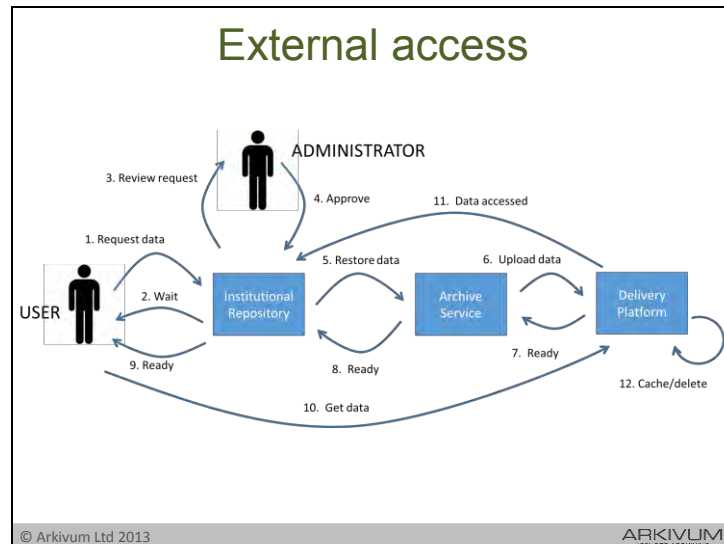
ARKIVUM

The need for a review/approve process falls into two main areas:

- The data set is large and time/money/cost is incurred to satisfy an access request.
- The data is private, confidential, embargoed or otherwise restricted.

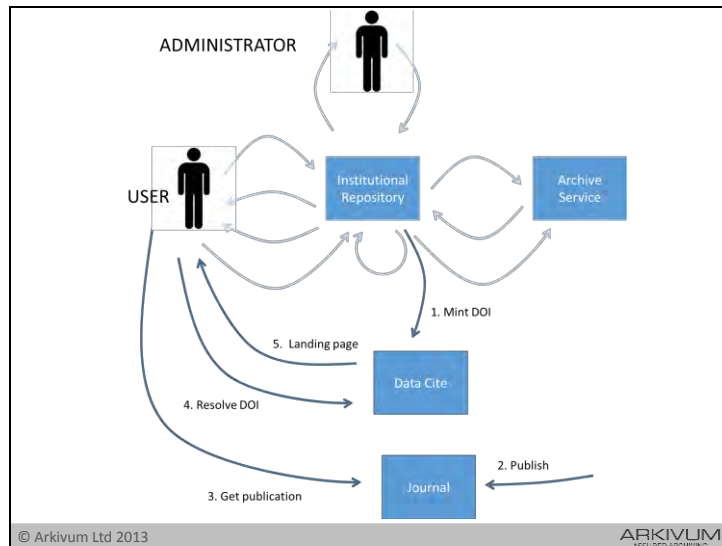
In either case, information on the person requesting the data will typically be needed, e.g. who they are, why they want to use the data, when and how.

Approval could also mean asking the creator/owner of the data – if they are still around. However, it is better to decouple on-going access from initial people who created data – this is a fundamental part of the OAIS model, although not always easy to achieve in practice.



As with deposit, it could make more sense for physical retrieval to take place direct from the archive, or by the archive restoring the data to a delivery platform, e.g. a ftp or webserver, which then provides access rather than the IR. This would make sense for large data sets. Or it could make sense where access requests will be frequent – for example by restoring data to a delivery platform somewhere more central on the JANET network rather than expecting the IR to cope with heavy access traffic and to do so over an institutions own JANET link.

This alters the order of steps in the workflow. The IR remains the gateway to access requests and keeps a record of all the accesses made.



Initial discovery of the research data might not be from the IR but instead through publications or portals where a DOI is used to resolve the data set to a specific location or landing page. For example, the datacite plugin to eprints allows easy minting of DOIs.

Asynchronous interactions

- Deposits and restores can take time
 - E.g. large datasets
- Review/approve process may be manual
 - Multiple stakeholders
- Request / ready protocols needed for users
 - Request forms
 - Notification mechanisms, e.g. email
- Deposit / restore protocols needed for systems
 - Plan, Do, Check, Act
 - APIs, callbacks, polling, events

© Arkivum Ltd 2013

ARKIVUM

Irrespective of whether data is accessed through the IR or direct from the archive, there will be the need to design and support asynchronous processes – both because of a review/approve process but also for large data sets that take time to retrieve and deliver to users.

In the case of interacting with users, then protocols for request/ready are needed along with supporting request forms and notification mechanisms when data is ready.

In the case of interactions between IR and archive, this is more automatable so there is direct system to system interaction to support, which means use of APIs with supporting asynchronous communication mechanisms, e.g. use of call backs, polling or event based messaging.

Interactions (2)

- Knowing stuff is still safe
 - Audit trails
 - Tests and validation
- Retention policies
 - Holds
 - Deletes
 - Reviews

© Arkivum Ltd 2013 ARKIVUM

Having looked at getting stuff in and out of the archive, now comes the issue of knowing it is still safe and also managing retention including being able to delete data if required by a policy, or 'holds' on deletion if this needs to be blocked.

Audit trails

- Essential in compliance applications
- Part of records management
- Digital preservation is an active process
- Good practice to record checks/interventions



Modular Requirements
for Records Systems

ISO 23081-1:2006
Information and documentation -- Records management
processes -- Metadata for records -- Part 1: Principles

ISO 15489-1:2001
Information and documentation -- Records
management -- Part 1: General

© Arkivum Ltd 2013ARKIVUM

Knowing that data remains secure and safe in the archive means audit trails are important to confirm that preservation is being done successfully and that the data has been properly secured with controlled access. This is particularly important given that preservation is an active process requiring a range of interventions (see previous webinars for more on this). Retention, deletion and holds should also be recorded as part of an audit trail.

Audit trails are essential in regulated environments, e.g. for clinical records, but their use is good practice anyway.

Audi trails: events

- Type of event (e.g. ingest, encryption, replication)
- When it was performed (i.e. timestamp)
- Who it was done for (e.g. user, admin, system)
- Where it took place (e.g. data centre, local)
- Why it was performed (e.g. security, access)
- Which data involved (e.g. file or set of files)
- What metadata was changed

© Arkivum Ltd 2013

ARKIVUM
Digital Preservation

Audit trails consist of events, i.e. things that happen, and changes to files or metadata as a result of these events.

Events use a 'who, when, why, where, what' type model. The 'why' bit is perhaps overlooked more than the other attributes of an event, but recording the reason for an event is really important for regulated environments to provide auditable justification of actions that can or do involve access to or changes of content.

Audit trails: metadata

- File name and file path
- Location of the file (e.g. particular data tape)
- Checksums
- Encryption keys
- Access restrictions or deletion
- Media/storage migrations
- Format migrations
- Integrity checks

© Arkivum Ltd 2013ARKIVUM

The metadata that describes data in the archive will typically be updated as a result of events, for example an integrity test might generate a checksum and this is added to the metadata, or a file is moved between storage servers as part of migration, which means its new physical storage attributes will be updated in the metadata.

The audit trail should record both the events that took place and the metadata changes that result.

Retention policies

- Review retention against requirements
 - E.g. DPA, EPSRC, Patient Records
 - Review/approve process with multiple stakeholders
- Stop deletion (hold)
 - E.g. during possible litigation or FOI
- Stop access
 - E.g. if retention period passed but no action
- Delete
 - Multiple copies in different locations
 - Multiple levels of deletion
 - Tightly controlled access to delete functionality
 - Audit trail for whole process

© Arkivum Ltd 2013

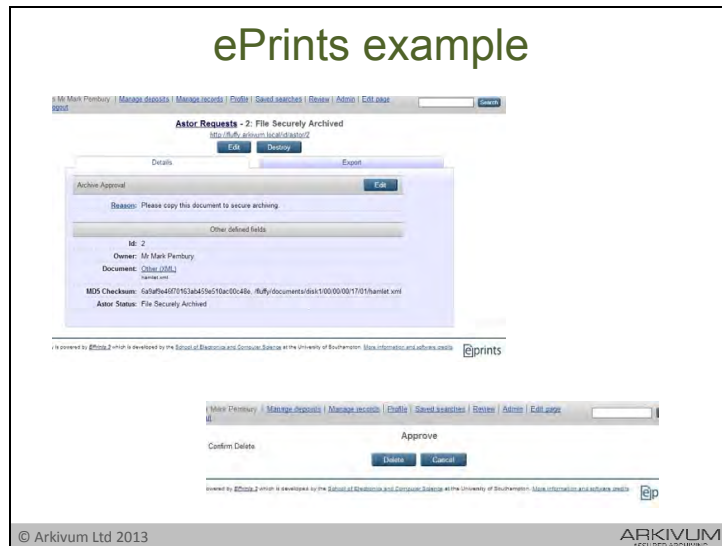
ARKIVUM

Retention policies say when a file should, or shouldn't be deleted. The IR will typically manage the policy and this can then be enforced by the archive, e.g. by making files WORM to prevent deletion. If a retention period has passed and data is up for review but no decision has been made, then an access block may be required to ensure no unauthorised access takes place – for example data held under DPA where the period for keeping the data has passed and hence there is no longer a justification to hold it which in turn means it should not be accessible.

When deleting data, it is important to remember that this may not be an instantaneous process. If the archive holds multiple copies of the data in multiple locations and some of these copies are offline then deleting all of them will take time. There may also be different levels of deletion to apply, e.g. initial crypto-delete by deleting the keys followed by physical deletion of the data itself.

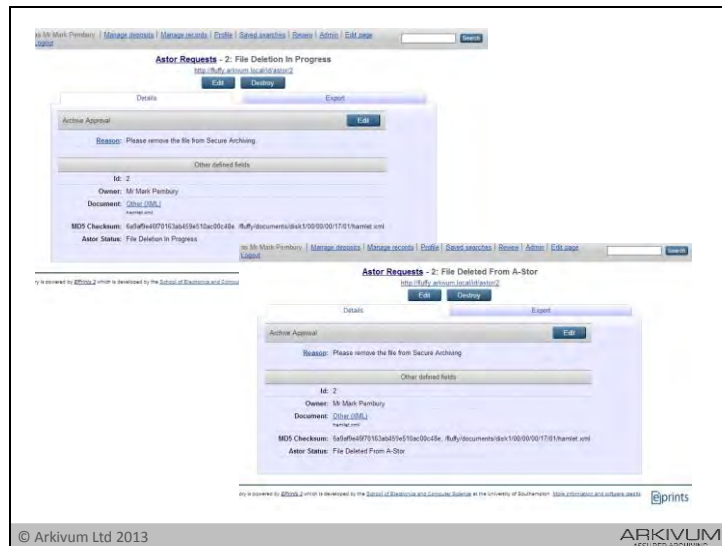
Given the primary function of the archive is for retention of data in a way that is safe and secure, any delete functionality that is implemented needs to have strong access control to stop accidental or deliberate misuse.

Slide 41



Arkivum and ePrints have worked together on an initial framework for retention management and data deletion. This screen shows the request/approve process for deleting data.

Slide 42



If a deletion is approved, then ePrints instructs the archive to remove the selected document. ePrints is then updated when deletion has been confirmed.



The slide is titled "Arkivum service" in a green font. Below the title is a bulleted list of four features: "Data archiving as a service", "Persistent names and paths", "Chain of custody and audit trails", and "Integration with ePrints and DSpace". At the bottom of the slide is a dark grey rectangular box containing the Arkivum logo (stylized "ARKIVUM" in blue and yellow with "ASSURED ARCHIVING" in small text below it) and the text "100% data integrity guarantee" followed by "Keep your data safe & secure forever". The footer of the slide includes "© Arkivum Ltd 2013" on the left and the "ARKIVUM" logo on the right.

Arkivum service

- Data archiving as a service
- Persistent names and paths
- Chain of custody and audit trails
- Integration with ePrints and DSpace

ARKIVUM
ASSURED ARCHIVING

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

© Arkivum Ltd 2013

ARKIVUM

Having looked at some of the specific interactions between Arkivum's archive service and ePrints, which is still in early days development, a few words about the Arkivum service in the context of IR integration. There's lots more information about the general Arkivum service in previous webinars.

Arkivum provides data archiving as a service. The data stored in the service has persistent names and paths meaning that DOIs associated with it won't suffer from broken links over time. Arkivum has customers in healthcare and lifesciences so we support audit trails and chain of custody.

The webinar has shown some of the initial work we've done on ePrints integration with Southampton. We're developing partnerships with other Universities so there is more integration support still to come. Likewise, we will be starting work on Dspace integration over the next couple of months.

Summary

- Archiving is an effective strategy for retention and access to research data
- Institutional Repositories have a pivotal role
- Deposit and access: mediated or direct?
- Integration is not just adding more storage
- Need processes, checks and balances

The webinar has hopefully shown the benefits of using archiving as part of a DRM strategy and the pivotal role that the IR including as an intermediary between the archive and producers and consumers of research data. This mediation model does require some thought for large data sets or where external access needs to be provided through delivery platforms or landing pages, in which case it may make more sense to provide direct access to the archive (properly secured of course) or use a model where the archive can push data to other systems for subsequent access. And in all of this is the need for proper processes, checks and balances - including chain of custody, retention policies and audit trails to ensure that data is safe and secure for the long-term – and this can be verified through the IR.