Research Data Network

Purpose of workshop

The network: “Place for people to communicate, share and pick up ideas around research data”

Rachel Bruce – aim of today – an introduction to the rdss and start sharing information around different issues; format mainly breakout sessions and discussion.
Idea for RDN – Jisc was tasked to take further work around research data management, in response to the EPSRC mandate. People said they missed some of the human network around the RD management programme; now seem to be the right time to form the people network around the RDSS pilot. We don’t intend it to be just around Jisc work, but ideas and solutions from everyone. Enable people to find practical solutions.
Hosted quarterly. Pilot institutions are volunteering and kindly offering their universities as venue. Next – Cambridge in September.
Another space for sharing information – online space to collaborate: research data network collaboration site. http://researchdata.network
Research at Risk – highest priority was a research data shared service and addressing preservation.

Welcome from Cardiff

Neil Penry
An intro to Cardiff, the wettest city in the UK.
Cardiff University founded in 1883. Member of the Russell Group. £0.5bn of research income.

Programme - research data information management at Cardiff:
-business case in december 2013 to the University Executive Board
-aim: set up CRIS, public portal to show research etc
-aim for RIM: CRIS with automatic feeds, capture and reuse research information so researchers spend less time doing admin, make REF 2020 a little bit easier and less rushed at the end, want to know more about our research and the linkages between different groups
-so far - procured CONVERIS, implemented research data repository with metadata, 80 datasets in there at the moment; 2 months ago went live with impact recording and so far 400 recordings on the system; have an interim RDS and using it to store the EPSRC datasets as well;
-in development: Eprints, but eventually will use Converis, also for all the funding application work

-the reason we bought Converis is because it has the ability to develop own functionality, also ORCID integration

RDM aims and future implementations:
-storage for live data
-help researchers decide what to keep
-automated movement to archive storage

Main RDM issues:
-shifting tech landscape and move to cloud computing
-the costs of the end provision
-how to see the whole thing to researchers
-how to fund the service
-integrations and interoperability

RB - picked up on the point of bringing I into RDIM

Update on research data shared service

Catherine Grout - head of change research, Jisc
The reason we are doing this is because there is not one or a few products in the market place that do this. There are various different components and platforms available, but mostly they don't talk to each other. At the moment we are working on a high level business case for research data and why it's required. The main thing is that RDSS has to be better and cheaper, otherwise it's not worth it.

At the moment it is much easier to go to PC world and buy some storage than going through the process of cleaning it up and adding the metadata.
Also, it is about improving the integrity of research.

Key requirement is preservation.

Goal - get the metadata right. We have started this process, we have the spine of the metadata.

Features: easy to use and implement, as well as interoperable.

Where are we now, an interesting journey up to this point, gathering requirements and designing the service took some effort; we have gone through the procurement process - slides showing the suppliers on the framework.
Next step - **suppliers have to articulate clearly what they can offer.**

Jisc portfolio for Research at Risk - we aim to align all projects.

There are other key areas in research data shared service that need more work so we are engaging with consultants for technical support. Slide showing the list of consultants and the key areas they are working on.

The list of pilots - goal was to draw a wide set of institutions at different maturity levels and with different requirements.

A lot of the suppliers never worked together before or with some institutions so we are hoping that the level of understanding for requirements will increase.

As we build our understanding and get in place concrete ideas we will start to assemble and organise the outputs and the sharing of information and use cases from the service.

Marta Teperek, University of Cambridge - why we wanted to be part of the pilot
-we have a repository for research data and about 500 datasets, but we have some problems internally
-it is the people network, share problems and offer/give support
-preservation for us is important, especially because we cover a wide range of subject and a huge number of different file formats; hope we can work jointly
-another big problem is big data; in theory we are able to help researchers with big files, but when we tell them about the charges, they are not that interested anymore; thinking about the workflow, how to get the data from the researcher, how to break it down and make the process scalable
-another big issue is data that has to be restricted for some reason - commercial data, personal data - who will manage the access request, who will be responding, are we confident that the technical aspects of our repository is robust enough to host this kind of data
People network is really valuable; the cost and time efficiency since we don't have to go through the same process at each institution.

Rachel Bruce on research data spring and the 'filling the preservation gap ' project by Hull and York University.

Q. (Nottingham Trent Uni) We are interested where it may go, your roadmap and where in the future you expect it to be, and when our institutions will be able to take it on?
RB: around about July 2018 - a production service, the data changes slightly because of contracts/signatures. We are aware that in some way this is not fast enough. For example the

market research, we are trying to do this early on and throughout the process; speaking with universities and understanding where you are.

Q. Is there a risk assessment for if the business case doesn't succeed?
RB: Priority - it is the number one priority in the research and development portfolio at Jisc along with Learning Analytics. Also UUK see it as important, so it has a high level of buy in, which helps with some risks. Also, if things don't work out, we are looking at migration, so institutions can migrate their data into other systems. Possibilities of sharing platforms. With the business case, there is wide applicability, so there are some aspects which could fit into the core offer and some more niche.

(1) UK Research Data Discovery Service

Chris Brown and Dom Fripp, Jisc
Landscape - there are a number of systems developed to expose data of all sorts, but not many national data discovery service, just in Australia and one being developed in Canada at the moment.
CB going over the history of the project - Phase 1 from October 2013 and Mar 2014, DCC and UK Data Archive involved.
In Phase 2 it's a Jisc led project. Now engaging with HEIs and UKDA dealing with engagement with the data centres. Project ends in September and we will have a test service in place.
Introduction to the team: Catherine Grout, Christopher Brown, Dom Fripp, Ade Stevenson, Veerle van den Eynden (UKDS), Diana Sisu (HEI Engagement, DCC)

Stakeholders - the list of HEIs and Data Centres.
Governance structure of the project:
-user group
-technical metadata group
-research group - because they are one of the key users, and so we wanted to get their input, and they are becoming more involved as testers as the project is progressing

Benefits, or why are we doing this?
-having the data in a central registry it would increase the visibility and transparency to the data, to remove the silos, it encourages sharing and validates research
-could also indirectly improve the RDM within institutions
-important layer in the research data infrastructure and fits with the shared service
-reduces barriers to research, if you find the data you can analyse it
-can also find out who is working in different areas

What is it for?

We gathered user stories in a workshop and did a MoSCoW prioritisation.
https://rdds.jiscinvolve.org/wp/2015/05/08/initial-workshop/
Everything we do is in google docs and done in quite an open process, so all documents and requirements are available on the blog.
We also used CKAN because it is more open. All kept openly so that if later on we decided it wasn't a good idea, we could go back and use something else.

Alpha site: http://ckan.data.alpha.jisc.ac.uk/
Publicly available and has been tweeted, there is a feedback link on it as well, so feel free to comment. All the data that has been harvested is of different quality.
We have now started to develop a business case

Current issues:
-quality - both content and completeness; we encourage people to put in as much as possible; the more information you have the easier it is to discover it
-have a core schema to allow consistency, but also additional fields
-licensing and copyright issues
-it's open so we cannot really do a login to give people analytics on who has been using their data; it's one of the issues being discussed
-updates and harvesting the metadata needs to happen a bit more often

Quick demo.

Metadata

Dom Fripp (senior curation metadata developer)
First looked at the criteria
-user requirements
-should map onto CKAN
-existing schema
-something flexible

The portals Dom looked at: EUDAT, ANDS, sparkle layer on European Data Portal, Data Citation Index, DataCite schema, OpenAIRE for alignment, ETSIN a small discovery service
Looked at what mandatory fields they had and which ones were in common; first semantically assigned, but then checked manually that they were equivalent.

Principles: findable, accessible, interoperable, reusable
Mapping OAI-PMH formats (available on https://rdds.jiscinvolve.org)

There is a demand for some or all of it to go to the shared service, some harmonisation and version control will be required.

Aim to make the schema richer to use within repositories; also looking at ways at harvesting across multiple formats to be able to collect it from different places, starting work on this on Monday with a few institutions.

Other institutions are happy to join the service, though we aren't currently going to harvest from any new institutions.

Q. Why is metadata licensed?
DF: for example visual arts, you want to provide an artistic description, from their perspective it can also be part of the art work, such as the title of the work of art or media used etc; this description in itself may be a piece of work thus cannot always be CC0. If you cannot share metadata it makes discovery redundant.
CB: report done by UKDA on these issues. If anyone wants to be involved in this project, welcome to get in touch if you want the service to harvest the data.

Q. Only harvesting metadata?
CB: yes only harvesting metadata and the link back to the data itself which stays within the institution. The data has to be available, if there is a login behind it, you obviously wouldn't have access to the source.

DF - if anyone has an idea about the OAI-PMH, get in touch, I would like to talk to you even if you are not involved in the projects, i would love to know about this area as it is a technical challenge

(3) Stop Press: Should Embargo conditions apply to metadata?

Sarah Middle, Repository Manager, University of Cambridge

Warwick, Surrey, Lincoln, DCC

Unanticipated issue, but comes up fairly often.

What's an embargo:

- pre-publication/ press - pre access for journalists to arrange interviews/ make a summary
- Post publication - add something to repository but not public release
- Restricted: to subscriptions
- No publication - security - commercially, personally sensitive, never released to public domain

Publications and dataset embargoes
Publications often easy as publishers have a policy
Datasets trickier as no policies on the whole. About 50% of researchers apply an embargo

Confusion about what constitutes an embargo - different between releasing the dataset and its descriptive metadata
Concerns that releasing metadata prior to publication could result in sanctions - but not sure of any actual examples of this happening.

Real researcher correspondence - confusion over abstracts being available; seeing data submission and paper publication as intrinsically linked.
“I have just seen that our paper...has been posted online on the repository of the UNiversity… Could you please take this down immediately? THis paper is under strict embargo from Science Magazine… and a breach of the embargo such as this can cause the paper not to be published in the magazine.” (Science)
THis was actually a case where the metadata/ dataset was available - in fact the paper was not. However, from repository there is no indication that the paper is not available.

Cambridge has done some research into publisher policies - majority of publishers have little or no information about releasing metadata. Embargo policies can be open to interpretation, i..e publishers can prohibit authors from discussing the fact that their paper has been accepted for publication.

“Nature journals will not prohibit metadata release (author, title, abstract, journal) for accepted articles ass per the HEFCE policy”

Some don’t prohibit release of metadata, but don’t encourage it either. - but this would result in less press interest
Science would prefer a cautious option of keeping the metadata in a “dark archive” until publication!! But are considering this as an issue (esp in UK due to HEFCE policy)

Cambridge strategy
Advise researchers on researcher policy
Re-assure that pre-publication metadata release is not a problem
Some other institutions are much more cautious and embargo all metadata records then manually release on publication.

Observations on way of working
Goes against Open Data principles
Creates extra step in workflow
Potential for delay in making dataset available following publication, resulting in paper temporarily containing broken links to dataset
Breaches requirement of DOIs to have a persistent landing page - will be minting DOIs for all datasets, embargos result in breaches

Solutions?
Didn't realise it was an issue, so now need better communication with researchers.
Need resolution between what institutions and publishers want - publishers want their embargo periods - and don't nec want the institutions to make the metadata available as can affect their embargo
Minimal metadata would be title of dataset and authors (nothing to do with the paper) - but need to be discoverable!!
Open Scholarship initiative is looking at publication embargoes (Neil Jacobs is Jisc person) - concern in community that publishers are trying to impose their view of the world. Steven Hill from Funding Council also engaged with initiative.

Collating case studies? Jisc role? Where there have been ACTUAL issues and not just perceived issues? (perhaps as part of the network? Also potentially belongs to work in Neil Jacobs team? They could tackle it, link to SHERPA services, RIOXX profile and metadata)

Metadata
Access to a dataset before publication of an article, the dataset does not have to be connected to the article? Existence of data should be available as early as possible. Certainly some concrete examples from theses being published previously, and then not allowing publications based on the thesis. (RDM list)

Is data directly supporting an article - versus publishing data that is a broader set
Cases causing concerns at Cambridge have been about datasets that directly support articles.

Are data publications tied to publisher requirements? - impetus behind publication of datasets is often (usually) in support of publication of a paper currently.

Thoughts for future:

- Want to encourage publishers to be clear and transparent about their policy
- Want to collect real case studies about actual difficulties (in interpretation and in actual obstruction of publication).

Sharing - is this an issue that's owned by the sector? SM would prefer that initially case studies were not shared publicly, but future advice should be developed to support researchers.
Possible that issue could be discussed in current Jisc fora.

Definitive statements about metadata from publishers could feed into a SHERPA Romeo search.

Q: issue of manually changing status of metadata based on publication date - if journal website supplied publications dates couldn't this be automated?
In DSpace can attach an embargo period based on the publication date on the actual data files, so that files become available only after publication. However, to make metadata invisible, the whole record has to be moved to a restricted collection - there isn't a way to automate its transfer to a publically available space

RB: From a Jisc perspective we will follow this up with Neil Jacob's team ; also some follow up with DKs project.

(4) Demo OwnCloud and Archivematica

Matthew Addis

DIgital preservation is about value on the long term. Costs of doing this are lower than the benefits. It's a no brainer, but usually not an immediate priority for institutions.

Article: parsimonious preservation… by Tim Collins:
-need to know what you have (file format, software to read the data now and in the future)
-keep things safe so it won't get lost

Archivematica - digital preservation tool that enables many of these things, open source from Artefactual in Canada, they develop and support Achivematica. There is an international user base and a UK user group (about 40 or so people, including York and Hull), next meeting is 13 September 2016.

University of York and Hull have done some work via Research Data Spring around Archivematica and how it can be applied to research data. Reports available on figshare.

Will demo archivematica but in a way it can be provided as a service: chain of custody, preservation actions, storing etc.

OwnCloud - like dropbox but you can run it on your own infrastructure.

Demo:
Local folder, ownCloud folder and the web interface of archivematica are open on Matthew's desktop. Bundling and zipping into a folder that will transfer them into archivematica.

Archivematica will identify the formats and make preservation decisions on those files.
You can ask archivematica to change the file format for you if you think one will be obsolete.

It is ready to go so if anyone wants to use it.
Q. APIs for archivematica?
MA - yes there are APIs, can automate submission to it and pull stuff from it.

Q. Dissemination?
MA - archivematica can also create bundle of files for public availability and dissemination, so would generate things like a thumbnail.

Q. Metadata?
MA - ??

(5) Business case and the costing model for storage volume requirements - a Royal Holloway case study

Frances Madden and Dave Cobb, Royal Holloway University

...Paul's notes go here….

Methodology – user types
Low, medium, high and massive categories of users.
Offline access also required
Looked at bandwidth reqs particularly with machine-generated data. Hard to get accurate figures from depts. Relatively low bandwidth required in practice. Looked at solutions that provide collaborative environments. How to make best use of architecture? Cloud options always came out cheaper when looking at large storage reqs.
Worked with research office - looked at grant funding to see if correlation with use.
40% of research in institution not funded – based on TRAC return.
Challenging trying to justify information on data – 47 page report. Created full spread of options to consider with estimated risk.
250 research academics, 750 all research staff. Rises to 2100 for all research staff and students. Used user numbers for active storage and project numbers for archive.
For archive data looked at how it could be reduced. Range of costs based on what's archived for different categories of users.
Other costs - integrations of systems/tools, service levels, project costs vs on-going costs, cost recovery.
No resources in-house so budgeted for staff. Service levels also includes application support and advisory service.

Unlikely that these costs could be absorbed with current staff.
Pulled in resources from different teams to investigate costs. This has cost implications across projects.
Cost recovery - estimated 35% return over 4-5 year grant lifecycle (based on TRAC return).
Findings:
Balancing cost and control is a challenge.
Perceived risks of using Cloud but already used by half of researchers. Risk is organisations use of Cloud not Cloud itself. Can't say at this point which Cloud solution was chosen.
Data provision as it stands is not adequate.
Very time consuming process, involving many people and information from disparate sources.
Positive feedback when initially cautious.
Next steps:
Procure data catalogue and archive storage before Aug 2016
Procure active storage solution to roll out Sept 16 - Jan 17
Integrate planning tool with other systems
Develop service around the system to support it and researchers using it.
Still need a plan for analogue data storage - more of this than expected. Also, training for admin and end users on all new systems.

Q) How to get data back from Cloud if academic leaves? Other systems hook in and once access removed control of access to these files is transferred to someone else.
Looking at unlimited storage solution for active data. 2TB of storage for archive data per project but not expecting all using this amount. It would be a problem if they did

(6) Update on Journal Research Data Policy project

David Kernohan

Based on Jisc project JORD - to assess the viability of a registry service for journal data policies to complement the information made available via SHERPA services in future
Convened an expert group to support the project, and some in house analysis of policies was carried out.

http://doi.org/10.1629/uksg.28
"A high degree of subjectivity and interpretation"... had to be applied to reading journal data policies

Primary issue: points of differentiation
16 questions used for data collection, but did not offer a complete set of user requirements (and were based on subjective policy interpretations) - i.e. user has to read the policy!

Templates and guidance for journal publishers around research data policies. To revisit the idea of a policy registry, including an investigation of the potential for machine-readable licence based solutions.
Journal policies become data policies and not much consistency between them. If policies were machine-readable, then some of the subjectivity could be avoided

Moving towards a template
Data are {peer reviewed along with the paper in questions/ peer reviewed separately/ not peer reviewed}
Data should be {submitted along with the paper/ submitted on request/ openly available at the time of submission etc}…

Short term continue conversation with stakeholders including publishers (project reference group, which includes RCUK and Wellcome (funders included as funder data policies may not link up with publisher data policies - idea of convergence has been raised!) and elsewhere)
Linked development of guidance and support in order to influence practice
Integration with international efforts in policies to support open access and open data (developments in Spain, Australia and elsewhere)

Jisc would like ot encourage debate amongst publishers around standardisation. There needs to be further work around disciplinary practice differences - link with some bio-medical work going on at Oxford - practice is being developed. Journal policy differences may be linked to domain policies, so linking may be possible. Publishers are not aware of other publishers practice. Jisc will be looking at alternate means of developing a policy registry.

Martin Donnelly: interested in policy from funder point of view - not just machine readable, but machine actionable policies - issues - terminology different and until there is an agreed taxonomy that won't be overcome - linguistic differences also come to play. CASRAI has a dictionary of terms - Jisc pilot was first international - UK community could be a place for a template/ common terms to be discussed, FOSTER also has set out some terminology

Another level of complexity at domain level - e.g. meanings of ontology different in different subjects!
Field level also introduces norms as far as type, quality, quantity of data required.

Marta: worth remembering that repositories, and therefore institutions are the publishers of the data.
RB: don't want to perpetuate mistakes made by paper publications
(Lincoln) - but we are still using the same language - issue with "publisher" for datasets.
RB: PLOS policy is a good starting point
DK: but even PLOS has requirements, positions and recommendations… which could be subjective as to how they are implemented locally

Marta: share publishers reactions - are they open to conversation?
DK: at last expert meeting were keen to press Jisc to develop support for journal data policies. Would have expected them to be further on than they are.

RB: How to generate a sector discussion about what's needed - it's very complex but might be useful to have a discussion about what Jisc can MOST effectively offer the community
DK: Impact from people in the community who have actually had to read the policies in order to support their researchers locally (data repository managers etc). Would be great to amplify those voices from institutions and use their knowledge.
(Lincoln) - problems but also opportunities - less existing practice to mirror - eg shouldn't use licences that are appropriate to paper publications for data - should choose a data licence.
Marta: - but we need to act fast - from the community perhaps a dedicated workshop to work on the licence/ template (like the RDSS workshop). Rem that researchers like FAQs, a simple explanation of policy documents.
"Policies are not written to be read, they are written to be cited"! We could make them more readable.
Lots of people are converging on a policy framework (e.g. PASTEUR)
Likely to be proof of concept in the first instance - without going to RDA would only be useful for UK - would need to be taken internationally

Rem that from point of view of publishers there will need to be a commercial angle

DK: ANDS - at point of talking to publishers now.
RB: ANDS, Jisc, SURF etc were taking forward

Relationships between publications and underlying data need to be considered, as data can be used for a variety of purposes, and can underpin more than one piece of work.

[Community workshop]

(7) DataVault demo

Stuart Lewis and Robin Taylor

http://pollev.com/rdmpoll/

Demo site is public, will include link.

Stuart explains how RDM works at Edinburgh - Data Management support (planning + active data infrastructure + data stewardship)
-ownCloud
-PURE CRIS

DataVault fits in the data stewardship area. Data repository is great, but you want to share only a small collection on the repository, so where do you keep the rest. Came up with the criteria:
-long-term archival storage
-multiple copies
-fast speed

A few analogies - the bank vault, if you want to store something securely, you fill in a form for metadata and then you come 5-10 years later, it should be still in there. If you put your grandmother's ring in, you don't expect the bank would have recast it into a more modern version. So the idea here is that we don't do much preservation.

Where does it sit - between the active data and archival storage.
DataVault is not a storage system, it's a platform. You still need to provide archival storage and active storage to your researchers. DataVault just transfers the data from one to another in a user friendly way.

It doesnt undertake active preservation activities like normalisation.

Use cases:
-golden copy data
-finalised data that can't be openly shared
-everything else

Good level of demand.

Phase 1: proof of concept
Phase 2: introduced the concept of users, admin dashboard, integrated with dropbox, and amazon glacier, introduced notion of groups
Phase 3: tidying up the last big bits, Robin is working on Shibboleth for authentication, working on retention policies, and CRIS integration

Fortnightly sprints, just coming to the end of Sprint 3.
It's all open source, and it's on github.

Demo: http://demo.datavaultplatform.org
To login user1 (password1) to user30 (password30)

At Edinburgh - we see use cases for both Archivematica and DataVault, we are going to use both, but it's about where to prioritise effort and what researchers would be more comfortable.

Q. Metadata
SL -We delegate metadata to the CRIS.

Q. CRIS integration rigid?
SL - all configurable

Q. Archivematica vs DataVault?
MA - they are not mutually exclusive and there are ways to integrate, though it might slow things down.

Q. Reason for using authentication system?
SL - for access to storage.
RT - typically would login with Shibboleth so you know who the person is and which people's data you can access etc.

Q. Have you worked out the funding for the other end?
SL - took us 1 year to figure out the charges and allocations for storage; this is going to be harder, there will have to be some free allocation but not too much so that you as a researcher are lazy and just dump everything. Don't yet know what's going to cost. It has to be a balance, cheap enough that it encourages people to use this as opposed to desk storage.

Q. Technology wise
SL - tiered storage GPAS??

Q. What will you do about the management of it?
SL - one of the metadata fields is the school you are in, every school has a data support manager who can login and see what all researchers in the school are doing and if researchers left the school etc.

Q. Can you delete data?
SL - no. We had a question, make it flexible first month in case there are mistakes, but we haven't done that.

DataVault Implementation workshop in Central London - 29 June 30 spaces for 15 institutions, will send out booking next week, talk to Stuart Lewis.

RCUK new grant submission system, Charlie Dormer, RCUK

Intro from Catherine Grout. RCUK process to introduce a new awards system.
Building a new system but not replacing ResearchFish.
Asking people what they need/want from the service.

Grants Funding Programme
Creating digital services supporting whole grant process. Until March 2017 to develop a basic product. Requirements coming in that won't be in the initial version. Turning Je-S off by 2018.
Replacing current system with a cost effective, flexible, interoperable platform
Should improve efficiency with improved experience. Reduce cost of giving out grants.

Ensure have data and MI to make smarter decisions.
Taking an Agile approach to development.
Simple functionality first - March 2017
Joint running with Je-S and new service for ~1 year.
Focussed on user need to develop iterative prototypes.
Current system - Siebel - large cost in training employees and not very flexible.

Very early look of prototype - 1 single page.
Using gov.uk design patterns.
End product will look less like gov.uk site.
Getting guidance into one place rather than four.
Simpler for new applicants.
Reduce re-input of data.
Questions will be asking for exactly what's required.
Doing away with attaching a PDF: bring everything they can to web forms.
Dashboard - work in progress. Have application, peer reviewing, writing.
Front end mirrors what is seen in the backend - not like the current system.

Outcomes from Jisc interoperability workshop last week.
User research shown strong desire to reduce re-keying and duplication of data.
Initial outcomes - people and org data (build on current work on ORCID), research costings (APIs between common costing tools and grants service + csv upload), equipment data, offer letter, student data, spend reporting, outcome reporting.
Ranges from easier to implement to ambitious.
Potential barriers - diversity of RO systems but same costing tool used by half at workshop; incomplete uptake of ORCID and standards implementation; implementation time; confidentiality, security.

Interested in further user research. If interested in taking part or vendors looking at interoperability get in touch. Charlie.dormer@digital.bis.gov.uk

How much are RC processes being harmonised? Lots of work on harmonisation, working closely with policy group. Not accommodating multiple ways of working in one system. RDM plans probably need to store more. Plan will be part of the form not an appendix. Have to investigate how it fits in with DMP Online.
Will guidance be accessible to services and not just applicants? Landing page is pre-sign-in so general/specific guidance will be available.
Any thought to interoperability with iris, trial-registration systems? Not at the moment.
Can't mandate one way of doing pre-application.
Looking at ORCID and RC SSO.
Needs to be optimised for mobile.

March 2017 - working end-to-end research grant system. Building functionality till replaces J-eS.

(8) Bristol case study on managing sensitive data

Kellie Snow (University of Bristol)

Data.bris went live in July 2014, was for open data only at that time, but there were already reports of requests for some data to be access controlled/restricted. Studies wanted to share but did not know how. Lots of informal processes going on and copying of documentation that weren't appropriate for each study.

Initial thoughts for restricted access:

What would any restricted access levels be?
How should they be applied
How would it fit with uni ethical policies
How could the service objectively decide who to grant access to?
What would be the processes that needed to be in place?
What are the technical challenges?
Did the service have capacity to deliver this?

First steps

Issue raised by Research Governance team
Meetings and discussions took place over several months
November 2014 - report put forward by the Research Data Service and Governance to Uni ethics committee: suggested levels of data access / Request to establish a Task and Finish Group (T&F)

Data Access T&F group
Established to investigate feasibility of a data access committee (DAC) as recommended by EAGDA report
Representation from IT security, governance, secretaries office (DPA/FOI), senior academics
Met 3 times - look at scope / data access agreements (drawn up by external lawyers) / process for controlled datasets
Many of the proposed procedures were based on the EAGDA recommendations and UKDA advice

October 2015 T&F produced a set of recommendations

- Access levels embedded in ethics application process

- Processes for handling restricted and controlled datasets piloted
- Permanent DAC formed
- Further consideration around retention of copies or requested data
- Consideration of data access levels at contracts stage (commercial research)

Passing through Uni Bristol committees - DAC soon to be approved (one more group to check)

Established Processes

Access levels
Open
Restricted - bona fide researchers. Staff check credentials
Controlled - requests evaluated by DAC, e.g. olde studies where permissions for sharing had not been granted or sought.
Closed - data not available for sharing (except regulators - requests handled by Information Rights Officer) because of ethics, IPR, etc.

How does this all work?

At start, researcher identifies likely access levels that will be used. Outlined in DMP. Allocated as part of the ethics application. Included in contracts with external partners (where necessary)

Data deposit stage, researcher assigns access level in the metadata record / verified by the RDS. DOI landing page repository record directs 3rd party to the RDS to request access form. Dataset itself is stored either within RDSF or with research group.

When request is received…

Restricted: access request form - background checks - decision if data can be released - send data access agreement to applicant - dataset under 2GB sent via FLUFF, over 2GB transferred over server.

Controlled: access request form received and acknowledged - conduct background checks - forward application to data steward / PI of project for suitability - DAC convened - send data access agreement to applicant - dataset under 2GB sent via FLUFF, over 2GB transferred over server.

Unforeseen challenges:
- Procedure for checking if applicants are bona fide - what does this mean? What can we check?
- Procedure for retaining data access agreements
- Storage location of datasets

- Supplying large datasets
- Researchers with controversial papers

Advice for other institutions

Never underestimate the time involved in getting agreement across the institution
You're unlikely to have covered every eventuality - be prepared to be reactive
Process will allow you to build better relationships with other university services
Your researchers will be grateful!

Future Plans

Develop Software system to deal with requests process
Develop an Online request form
Data access agreements repurposed by existing studies
Further procedures around commercial data (contract issues, IP)
Further knowledge and procedures around clinical data (esp. NHS) - where do sharing responsibilities lie?

Q: about sensitive large datasets - A: they can happen!
Q: how many requests do you receive and how long do they take to process?
A: Aim to do the whole process within 20 working days
Q: who is responsible to look after the data long term, if sensitive data is changing hands
A: this is addressed in the data access agreement at UoB
Q: What do you ask if they are bona fide?
A: Essentially, have to have an email address affiliated to institution so they can be checked against the institution. Also ask if they have a signatory within institution to sign off the data access agreement (e.g. IT security manager at UoB) - ask what they are planning to do with it?
Q: How is FLUFF secure? A: it is encrypted
Q: Is there a level between open and restricted? A: No. It was discussed but a lot of the levels were influenced by what the UKDA were doing.
Q: When depositors give something, are they allowed to say don't give it to these people / companies? A: would convene DAC to deal with that.
Q: in cases where the PI leaves institution, is there a handover procedure to another PI? A: in UoB RDM policy, there is a clause about having a named successor. (can defer to head of school)

>> Other institutions getting asked for secure sharing? SGUL / UCL /

(9) Demonstration of 4C Cost Comparison Tool

Most folks had heard of the 4C tool
Output from EC project to clarify costs of Curation (Excellent rating)
Curation costs exchange, cost comparison tool http://www.curationexchange.org/ to allow you to compare costs with peers
Focuses on the institutional repository, but is it suitable for the whole lifecycle?
Are you using CCT? If not, why not?
Comparing own costs with prior year costs is a compelling use case
Many people aren't submitting costs publically, due to commercial sensitivity
Create a cost set in the tool and:

- Map asset types
- Map to activities
- Map to purchases and staff

This allows for costs to be relevant to RDM and not simple finance line items
Entering type of institution allows for comparison against peers
There are generic asset types in the system, as well as cost units that can be compared
Activities mapping considers OAIS model

- Pre-ingest
- Ingest
- Archival Storage
- Access

Could be challenging to break down pre-ingest for RDM
Purchases include hardware, software etc. alongside staffing cost.

Paul Demo:

Can assign scope of data within organisation
Comparisons can be broken down by yesr and sliced by asset type and purchases and staff
You can also compare against the world and against your peers based on size of data and type of organisation
Help system highlights areas as you go along.
Currency comparisons are carried out in Euros on an annual basis
Privacy settings for information sharing anonymous or non-anonymous sharing of data
Q. Anna C - what are people sharing?

- A. Vast majority are private, you can contact through the system.

Demoed cost for Storage, Analysis, Security
There are disclosure issues, which may prevent comparisons where there are too few returns by certain organisations.
180 datasets in the tool so far.
Is this suitable for entire RDM lifecycle as a whole?
Could it be made suitable?
James Wilson: Doesn't define what is RDM e.g active data storage platforms, writing DMP's, focus on long term preservation. Needs drop down list of functions within RDM. Guidelines for costs, we need a template.

- Might be disagreement on what the activities are
- Work towards consenus

Are you using CCT? If not why not?
Anna C - Needs to cover all of RDM. Are there any worked examples (anonymous)

- Needs one, especially RDM

Video demo is available https://vimeo.com/117789715
Have you asked institutions directly? Yes, but sometimes finance block data.
Effort vs benefit - case needs to be made. It is difficult to get the fgures.
Session at iPRES with DPC

(10) Demonstration of DMA Online.

Masud Khokhar & Hardy Schwamm, Lancaster University

Jisc funded project under RDSpring.
Looking for new partners.
Use PURE as a CRIS system. 80 datasets on data catalogue.

Idea of dashboard to create useful tool to create reports. Pulling information together across disparate systems to do this. How could this be aggregated to help RD Managers? Created the dashboard. Simple idea - easy to digest info and make it useful for different audiences. Presents simple info but can dive in for more detail. Flexible.
4 institutions providing information.
If sign-in (demo with test inst) shows institutional information.
Don't know how many DM plans are being written in the inst. Want to find out more about DM plans using DMP Online. Recommended tool for academics.
Click All DMPs - view more info aggregated from various sources. Shows project and funder information. Can update DMP status to verified, for example.
Projects without DMPs shown.
Archived datasets shown - use PURE or metadata points to o39ther sources. Shows dataset, funder, dataset PID, lead faculty, lead dept, project. Can filter information on faculty.
Reporting functionality in PURE not great. Much better in DMA Online.

Use case - can decisions be made on the information? Comply with RCUK policy, increase storage, etc.
Can export faculty information.
Can name and shame departments not complying with RCUK policies.
Detailed costing model behind the system to calculate storage costs.
For every project have estimated storage size requirements. Size plus where stored (local/cloud), gives expected cost. Expect this from DMP Online in the future but currently add to system. Shows what's asked for against what's actually used.
RCUK compliance more difficult to achieve. Data access statements hidden in publication (PDF).
How many DOIs minted through Datacite can be shown.
Will link to Archivematica to show archived data.

At stage where need input from other users. Do you find it useful? What other use cases are there?
Doing API development with DMP Online to extract data.
Project finishes in September but will be used in Lancaster so will continue and be sustainable. Flexible solution.
Currently in alpha. Should be beta by September. Work will continue. Not looking at it as a business but needs a sustainability model.

(11) Presentation on UKDS Secure Lab and sensitive data

Dr James Scott UKDS

UKDS

- Funded by ESRC to support researchers who need high quality social and economicdata/
- Single point of access (many different styles of dataset: llarge scale, microdata, qualitative)
- Around 7000 datasets

Secure labe holds sixty datasets - detailed microdata - data deemed so sensitive - same security model as VML (at ONS) HMRC datalab and ADRN
Access remotely from researchers institution (subject to approval) (no travelling required)
Nothing goes in or out of secure lab environment without being checked by support team first

Principles: Safe model: safe projects / safe people / safe data / safe setting / safe outputs = safe USE

DPA1998
Provides framework to ensure that personal info is handled properly

What is personal data? Related to living individual which makes it possible to identify them , or expression of opinion about the individual

Different definition: Need to incorporate matters with non-living people as it can affect living people

DPA1998 should only disclose if given consent to do so, and if legally required to do so.
What is sensitive personal data? DPA says information relating to race, ethnicity, politics, sexual life, mental and physical health, offences.

Any data can be sensitive to somebody (not good idea to generalise)
Research exemption: Section 33 provides limited exemptions of some data protection principles to be processed for research purposes.

<u>SDC - statistical disclosure control</u>

- Checks on outputs to make sure they aren't disclosive
- Manual process carried about by 2 staff members
- Two approaches: Rules based / principles based
- We take a principles based approach.

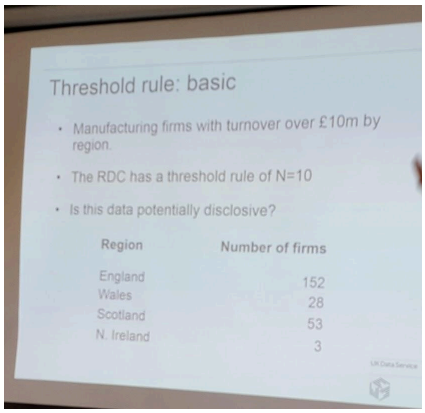
Outputs can vary considerably in size and length. Impossible to be prepared for every eventuality so always asked for some contextual analysis. Emphasis on researchers releasing as much as possible.

But 2 Rules of thumb employed

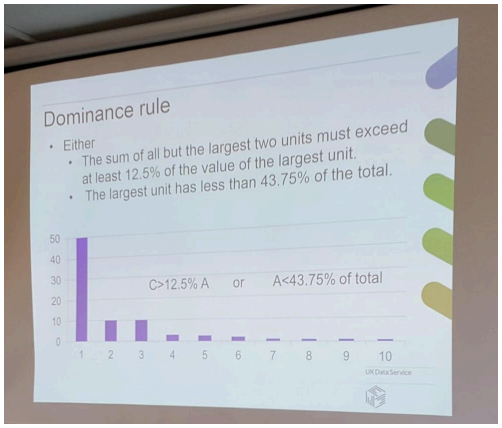
1. Threshold rule: no cells in a table should contain less than 10 observations (cf HMRC has 30 as their rule)
2. Dominance rule - no observation should dominate the data to a huge extent.

Why a threshold rule? Margin of error, 10 problematic for users but is high enough to make identification of individuals difficult.
Perception: small numbers can look unsafe even if not

Threshold Rule example



Dominance rule - Largest unit has to be less than 43.75% of the total.



Has to reflect the pledges that were made when the survey when it was conducted, doesn't matter if the info can be triangulated another way - have to honour the integrity of the consent process

Rules & Procedures

- Keep researchers within the law
- Comply with requirements of data owners
- Protect data subjects
- Ensure the continued operation of Secure Lab

Q: Is Securelab technology equivalent to working in a safe haven (like in IG toolkit)?
A: Using Citrix software to support access and restrict data going in or out.

(13) Demonstration of IRUS Data UK and metrics

David Kernohan, Jisc
Interactive session.
To investigate the feasibility and utility of various forms of usage stats concerning research data.
Download metrics and citation metrics a recurring request/need.
Brought together Expert Reference Group with experts in these fields.

Key activities.

1. Data download metrics - running as an alpha service. Soon to be beta. Reusing core IRUS-UK software and functionality. Collects file level download metrics from mixed use and data reps. Applying and developing COUNTER compliant download stats.
2. Data citation metrics - linking with global work on open citation metrics for data (FORCE11 DCIP, NISO, etc). Developing use cases. Give a clearer understanding of the theory of citation and data citation.

Project blog (https://rdmetrics.jiscinvolve.org/wp/) contains link to Cameron Neylon's discussion piece on data citation (http://repository.jisc.ac.uk/6399/).
IRUS portal available but in alpha stage.
Working actively with Figshare and Elsevier's PURE to access data by start of summer.
By end of summer will have full Jisc service.

Guidance, use cases and case studies outputs for data citation metrics work. Not developing software at this stage.

Can see date of when data download. Benchmarking against international standard.
Talking to Archaeology Data Service to integrate metrics. Measure of interest in the data based on downloads.

Qs from DK -
Any interest in becoming a test centre? Maybe in the future. Like to but use PURE so by summer.
What would you do with these stats? Show success stories. Often surprises academics about use. How to show re-use of data from downloads. Could use IP address to know it's Cardiff, for example, but can't know how used. Would need log-in but then not open and accessible. Privacy issues in tracking use.
Would data back up need for more staff, for example, within institution? Possibly. Shows bigger impact.
Anything people want to do with citation metrics? Eventually a separate system but long way off.
Is it only of interest to researchers? Metrics becoming more important. Yes, if needed for REF.
Needs more investigation to find out why data's been cited. Would require small number to be manageable.