Note: RDM readiness: meeting the EPSRC policy framework / SPEAKER-CONFIRMED VERSION
London. Fri 13 Feb 2015.
http://www.dcc.ac.uk/events/other-dcc-events/RDM-readiness

Ben Ryan, senior manager, research outcomes, EPSRC. 'The EPSRC Policy Framework'
https://prezi.com/kflylbtkcgvu/rdm-principles-and-expectations

- Reminder of RCUK principles; EPSRC Principles are almost word-for-word the same and the Expectations arise from them.
- Main messages are –
    - RD is a public good; should be made available responsibly.
    - RD has potential long-term value
    - What is kept should be discoverable and citable
    - EPSRC acknowledges legitimate concerns about what can be shared. That doesn't mean any restricted data shouldn't be discoverable: in a majority of cases it is anticipated that metadata can still be made discoverable. If there are constraints on access, then that should be clear in the identifying and descriptive information.
    - EPSRC understands discipline practice can vary. In some research communities there is a race to make data available, on which others can publish; but in others we recognise use of privileged access to data whilst publications are issued. This again should be planned and described.
    - Data users should cite their sources and abide by terms of access.
    - Sharing and looking after RD is part of the research process; and legitimate use of research budgets.
- These do not constitute a position unique to EPSRC: other councils – and some other funders – have very similar positions and expectations.
- We made it clear that the organisation has a role and responsibility for looking after RD: this is not just on the shoulders of researchers.
- Organisational responsibilities include establishment of infrastructure and processes to ensure:
    - Data which is selected for retention is retained for a minimum of 10 years after production or last use, whichever is later;
    - Effective data curation through full lifecycle;
    - Knowledge of publicly funded RD holdings is made widely available;
    - Discoverability is promoted and 3rd party access requests are recorded;
    - Where appropriate, notice and justification of access restrictions is made;
    - Awareness / use of relevant legislation e.g. FoI inc. specific exemption on research data;
    - Awareness and compliance with institutional RD policies;
    - Adequate RDM resource allocation is made e.g. from QR / grants.
- Researcher is responsible for:
    - Understanding and adhering to EPSRC principles and expectations for RDM. (It ought not to be a huge amount of extra effort, having got the data, to put it somewhere that people can see it.)
    - Complying with institutional policy;
    - Getting a DMP in place, even though EPSRC doesn't ask for it at bid submission stage;
    - Performing selection of data to be retained;
    - Ensuring appropriate agreements about data retention, rights, access, etc. are in place with any collaborators;
    - Ensuring the published research describes how to access supporting data (also required by the RCUK policy on OA). Not enough to put PI email address at bottom of publication. Email addresses volatile. Better to have persistent identifier pointing for example directly to the data, or to a landing page from which the data can be accessed, or to a data document describing the data and any constraints on access;
    - Being aware of and using relevant legislation and available exemptions as needed to justify withholding research data.
- Non academic partners also have responsibilities:
    - Understanding and accepting that publicly-funded research data is expected to be made freely and openly available with as few restrictions as possible;
    - Familiarity with relevant legal requirements such as the FoI regulations;

    - Need to ensure any proprietary data that is not to be made available following the research has clearly available reasons for restriction.
- EPSRC published further clarifications, including input from DCC, in October 2014.
- What's going to happen after 1 May 2015? EPSRC trying to strike balance between requirements and expectations about how business as usual will be operated in the institutions we support.
- Data preservation and RDM is still an evolving area. RCUK has said, with reference to OA, 'it's a journey, not an event': to a certain extent, this is the same. The proof of the pudding is, 'Is the data that is from this funded project available?' We need to see what this looks like, and that will inform our strategy in the future.
- After 1st May 2015, but before the summer break, EPSRC will be sending round a light touch self-assessment questionnaire, for the pro VC. It will say something along the lines of, 'The deadline is now past – how do you think you're doing at your institution?'
- After the summer break 2015, EPSRC will start a process of 'dip-sticking' published research papers to check availability of data underpinning published research. They want to understand what's going on at different institutions with regard to papers published after 1 May 2015.
- EPSRC will investigate any complaints that an organisation they fund is failing to ensure RD is managed in line with EPSRC expectations.
- If it is found that an institution is being deliberately un-engaged and obstructive, EPSRC will need to consider carefully whether they should continue to fund that institution.
- EPSRC aims to embed compliance checking through formal self assessment and follow up as part of regular 'dipstick' visits by the research councils Audit and Assurance Services Group (AASG).
- AASG approach to be 'business as usual' by May 2016? This is an open question – this is something the research councils need to consider together.
- Because a dataset can be reported as an outcome in Researchfish, Councils are currently discussing whether to *require* provision of persistent addresses and or identifier for all 'research datasets' – or documents describing them and how to access them – through Researchfish – if collected in this way and then exposed via GTR it would greatly enhance others' access to RC-funded data. However, this may not happen – it's just being discussed at the moment. Don't expect to see much happen on this imminently.
- Initially at least the main route to data is likely to be through the references provided in the publications themselves while institutions continue to make progress towards getting data catalogues in place.

Questions for Ben / comments / discussion:

- Q: Dipstick checking – just based on EPSRC research outputs, or on the overall capability of the institution? Ben: We'd be looking specifically at EPSRC-funded outputs.
- Q: Checking of data in published papers – published by 1st May or accepted for publication by 1st May? Ben: published rather than accepted. Framework has been out for 3.5 years now. Expectation of that data is that it will be available, and this has been clear for some time, so papers due out in May 2015 have likely been prepared during that period.
- Q: What if the HEI has policies in place but some prominent PIs won't engage? Ben: Good point. Researchers share responsibilities too. If institutions are making sure researchers have access to everything they need, that's all we can expect of the institution. There is the possibility of collecting at the end of project some sort of statement about access to data. If researcher is recalcitrant, we might say, 'don't come back to us for further funding', but that is at researcher level (rather than at the institution as a whole).
- Ben is sensing a desire to have one single policy across the seven councils. The councils are discussing this together. It's very difficult to do. The policy should be as non-prescriptive as possible, but some councils need to be more prescriptive. Councils do have the same principles, however. The differences in policy are mainly down to the differences in the nature of the data that tends to arise for each Council's funding.
- EPSRC does believe everyone should put DMPs in place. But we don't think they're primarily for the benefit of the funder, they're primarily for institutional and research process benefit.
- Q: We are being asked by researchers what makes them compliant. What is their minimum required data? Xml file? Raw data? Do you provide guidance on this? Ben: No. This would be a huge minefield if we tried to over-specify. If someone else is seeking to explore their work or reproduce their findings, no researcher should be in the position of saying, 'you'll have to take my word for it.' Data that is

required is the data necessary to validate, reproduce work, but obviously we want researchers to keep other data too if they believe it's likely to be useful in future. In the case of a very large raw dataset that has been substantially processed to produce a derived dataset, if the processing steps that have been gone through to achieve the derived data are thoroughly documented, that could be OK. You wouldn't need to keep the whole raw dataset as long others have access to the information they need to reproduce your results.

- Q: If you're going to use Researchfish, please let it be harmonised from the beginning. Ben: It would be across RCUK.
- Q: Re. the use of Researchfish, please can you use your influence to address the lack of upload facility for data. Ben: Researchfish is collecting information from about 30k researchers from about 200 institutions. A lot of the reports overlap. We have to crack the nut of unique identifiers for publications and for grants, accurately entered. Over next few months we'll be looking at this.
- Q: Re. persistent URLs for data – you mentioned you were considering this. I thought it was already a requirement to use digital object identifiers? Or have I misunderstood? Ben: We recommend that datasets are clearly identifiable through identifiers. If people are compliant, it shouldn't be a problem to tell us what those are. It's the use of Researchfish we're considering, not the use of permanent identifiers. Would it be a good idea to have section in Researchfish that prompts people to provide their unique data identifier(s)or to tick a box saying, 'this project has not resulted in any research data', which is possible in, for example, maths?
- Q: Is selection and appraisal activity including in the acceptable costs for RDM in project bids to EPSRC? Ben: Yes, if this is part of the activity that takes place during the project to support RDM. We want to encourage people to think about RDM effort before the end of the project. You don't need to put every RDM activity on its own line in the budget – we don't require that detail.
- Q: You mentioned that the PI email address is not sufficient for access to datasets. Ben: Yes. A robust identifier to data or to a document describing the limits of the access would be appropriate. Or at the least, an institutional, functional email address for a particular role or function rather than the email address for a named individual because people move around. Q: so the link for access doesn't have to be direct access to the data. Could be to someone who can send it. Ben: The data should be as freely available as possible subject to legitimate limits. If there is not a legitimate reason for keeping it private, make it available.

Jeremy Sharp, director of strategic technologies, Jisc: 'RDM readiness: supporting infrastructure services for RDM'
http://www.dcc.ac.uk/sites/default/files/documents/events/workshops/RDM-readiness/Supporting-infrastructures.pdf

- Infrastructure developed by Jisc includes: Datacentre and storage frameworks; access management services; network infrastructure (Janet).
- E-infrastructure is part of landscape along with open research data and big data analytics.
- Janet network: ~900 institutions connected across UK to very stable high-speed network. Supports big data and big research down to normal business of FE and HE. Provides world-class infrastructure.
- Attached institutions include universities but also facilities such as the Met office, and biomedical research facilities. They rely on this as a high-speed network that operates at high capacity. Most research-intensive universities fit in core of Janet infrastructure.
- Sits in a global context, part of European infrastructure – provides stable network to about 40 European research and education networks.
- BIS funding of £4M has been received to get Janet 'open and accessible' to industry.
- Provides industry access to university e-infrastructure facilities to facilitate further investment in science, engineering and tech with the active participation of business and industry. Modelled on Innovate UK competitions process. More information at www.ja.net/janet-reach.
- Working over last few years to develop frameworks for storage in different ways.
    - File sync and share;
    - Enhanced products avail offering EEA storage for data protection compliance;
    - User managed encryption;
    - Integrated federated access – end users can access services using institutionally-issued credentials;
    - Providers such as Box, Microsoft, and others are on the framework agreement.

- Single supplier framework with Arkivum: secure, cost effective data archiving service for research and education. 2 UK hosted data copies. 10-year framework to Dec 2023.
- Jeremy has a question for the audience: Is there a gap in national scale storage facilities for online or near online storage requirements, possibly complementing what is available at your institutions? Can we help put something in place?
- Range of compute solutions appropriate to research. E.g. institutional, regional, national, international. Architecture tailored to HPC or HTC.
- Shared data centre: £900K HEFCE investment. Anchor tenants: Crick, KCL, LSE, QMUL, Sanger, UCL – www.jisc.ac.uk/shared-data-centre. Difficulties in getting power and real estate in central London at that scale, so we worked with these named organisations to procure housing and an enterprise system further out of London. The data centre is connected to the Janet backbone at high speed: 100gb/s. A number of other universities are moving in.
- Provision also available for secure data management. UCL, Sanger and Crick are also collaborating on a biomedical project, Medlab, which has been recently installed. Framework with company called Infinity, available for all HE and research to use and also industry where there is a research collaboration.
- Requirements for a second data centre currently being gathered. North of England. Sensing requirements but don't yet have group of anchor tenants.
- The more datacentre is used, the pricing comes down across all tenants. Monthly recurrent cost is already decreasing as use increases.
- Assurance is provided through robust access control. Identity management provides authentication of the individual user. Group management and account allocation procedures provide authorisation for access to particular data holdings.
- Moonshot www.ja.net/moonshot: standardised to IETF. Single unifying technology to enable you to effectively manage and control access to range of web and non web services and applications, e.g. HPC, cloud infrastructure, grid computing and commonly deployed services such as email, firel store, remote access and instant messaging.
- Security: Janet has strong security team and trust network. Secure and reliable network for research and education. For example, working in project called Safeshare with biomedical research organisations including the Farr Institute, MRC Medical Bioinformatics initiative and ESRC administrative data centres. Network provides encrypted VPN infrastructure between organisations. Providing enhanced confidentiality and integrity per ISO27001 (Information Security Management). Aim: to result in more general services that we can add to Janet network services.
- Q: Is there a price structure available? A: For framework services, enquire through support desk. Use of the network is part of national infrastructure available through Jisc subscription.
- Q: Support for security? We have to comply with the Cyber Essentials Scheme. A: Building our capability. Good at reactive element of security. We want to increase our training and outreach. We'll be in a position very soon to do that.
- Q: Most universities connected to Janet. Are the framework services described today something that Jisc is going out to universities to offer, or are universities expected to contact Jisc? A: We are interested in hearing from universities interested in Safeshare. The ways that we can communicate with universities are now part of Comms since the reorganisation. The major change that will be noticeable is that the regional support centres have been reconfigured now into regional offices around UK to support FE and HE education and skills. This happened recently. We're still working out new processes.
- Jeremy posed question re. whether there is a gap in provision of e-infrastructure to which Jisc can response. A: If we can do something more cheaply through doing it nationally, everyone will want to take part.

Matthew Addis, Arkivum: 'Examples of combining research data archiving and access'.
http://www.dcc.ac.uk/sites/default/files/documents/events/workshops/RDM-readiness/Arkivum_EPSRC_Readiness.pdf

- We provide data archiving as a service with contractual guaranteed integrity. We have insurance backing. There is a data escrow facility so users are not locked into our service in any way. We are audited and certified to ISO27001.
- Presentation covers institutional and research points of view as, as Ben made clear, RDM is a shared responsibility.

- Looking at what we can offer that will provide integration with your RDM infrastructure to meet EPSRC requirements. Access to data doesn't have to be online and instant. There is a difference between what must be public and what can be private, and there's a difference between data being discoverable and providing direct and instant access to that data.
- Four quadrants of research data curation – slide from Edinburgh data blog: http://datablog.is.ed.ac.uk/2013/12/06/the-four-quadrants-of-research-data-curation-systems/
- A CRIS or IR can act as a gateway for doing RDM, helps with curation, helps to start preservation. Supports monitoring and metrics so you know when the data's being used and so whether it needs to be kept,
- but RDM begins and ends with researchers. There need to be benefits for researchers to use infrastructure. Some institutions addressing this in positive way with benefits to researchers including more citations, more collaboration, more funding. As well as less risk of rejection for funding. Easy to use interfaces for researchers to RDM infrastructure can also provide the benefit of keeping costs down and helps to make RDM part of day to day business – embedding RDM into research workflows, and reduce time needed to manage – and re-find – data.
- Case study: Loughborough. Figshare, Elements, Arkivum, DSpace.
    - Data and metadata goes to Figshare web interface. Some researchers were already using it and found it easy to use. Immediate benefit of getting it online. Goes to Figshare on Amazon-hosted cloud and gets a Datacite DOI (implication is that this is part of it being on Amazon-hosted cloud). Metadata and DOI to Elements. Replicated from there to IR. DOIs resolve to Figshare online. However, this doesn't necessarily work for large datasets. In that case: Figshare uploader but data goes to Arkivum.
    - Simple story to related to EPSRC: data is discoverable, automatic minting of DOIs. Being retained for 10 years. Upfront and pay as you go from Arkivum. Figshare lets you track who is accessing it. Adopted by researchers.
- Case study: ULCC. EPrints and archiving.
    - ULCC host EPrints and access to Arkivum so complete hosted solution.
    - Arkivum appliance to take the files – only removed from EPrints on confirmation that the data is safe and securely archived, in Arkivum service.
    - Access: request via EPrints. Small datasets: delivered immediately. Large datasets: review / approve process by editor, then released.
    - Managed access via EPrints. Open access: no license, no barriers. Through to embargoes, locked down.
    - Fits well with EPSRC clarifications.
    - Easy to adopt and use.
    - Work planned to handle large datasets – archive directly but adding links to the IR. V similar approach for DSpace and Pure as well. See the Research Data Spring for more info.
    - Fully hosted. RDM nursery. If you don't have your own infrastructure, you get this off the shelf, ready to go.
    - Provides one clear location for researchers to go. View / approve control process. Adding support for growing RDM.
- Case study: Aston. Arkivum
    - Institutional storage service. When ready for archiving, copy to Arkivum. When that data needs to be available, recall through Arkivum appliance.
    - Lowers costs.
    - Meets funder expectations.
    - Can build access on the top. We can add links in. if they have data now, they can get it in storage and if there's a request they can get it back again.
- There is a range of solutions. Include ease of use for the researcher, which can be pivotal.
- Q: Audited 6-monthly for ISO27001. Includes hosted service and onsite version.
- Q: Three data centres in the UK. Eduserve Swindon, Harrogate N3-conneted (NHS) and Janet, and Bristol.

Rachael Kotarski, British Library: 'DOIs for data: Using DataCite in the UK'
http://www.dcc.ac.uk/sites/default/files/documents/events/workshops/RDM-readiness/DOIs-data.pdf

- DataCite is mentioned in EPSRC guidance, expectation 5.
- There needs to be a link between the data location and the DOI. You're responsible for updating the location if you move the data.

- DataCite is standardised to ISO 26324: 2012 (Information and documentation - Digital object identifier system).
- Full members of DataCite can act as allocating agents, i.e. can mint DOIs.
- BL is one of 22 UK members of DataCite
- 'Should I be using DOIs or another identifier?' We have some questions to help you make that decision.
    - If you're not the data owner, you may need permission from the owner and creator.
    - Data should be maintained and useable for long term. We don't specify what 'long term' is.
    - Accessible doesn't mean instantly available openly – the user might need to register, obtain clearance for access, etc. If data can't be cited, why would it be stored?
    - Does the data have citation potential? Is it likely to be cited?
- Data centres are expected to provide:
    - mandatory CC0 metadata for its datasets;
    - 5 pieces of mandatory metadata;
    - publicly accessible landing page for each dataset;
    - and maintain working URLs for registered datasets;
    - curation and preservation policies. For whatever length of stewardship is decided upon.
- Form of DOIs: we place no restrictions on DOIs providing our criteria are met.
- We don't define data specifically. They can be applied to any digital object.
- Most users are also using the API for minting the DOI itself, and API for sending the data to DataCite. We have lots of plugins for different repository systems and bespoke repository builds, including Natural History Museum's revamped one with CKAN.
- Institutions sign up, get specific prefix. Institution specifies the suffix.
- Current members include Universities, NERC datacentres and other datacentres.
- DOIs are being applied to data files, grey literature, 'collections' (e.g. photographs, DNA structure, video, and micro-CT imaging data), crystal structure, poetry (e.g. Oxford). ADS has a bespoke system with integrated DOI minting.
- RK compared two datasets – each from crystallography and monograph – to show different amounts of metadata for each. Both examples are still compatible with DataCite.
- Rachel compared three landing pages already used with DataCite. Must have DOI and the mandatory metadata on the landing page. They can (and do) look different.
- Curation and preservation policies: point us to what you already have.
- Suffixes: there are different approaches. Bristol: long complex suffixes: they may be to encourage the user to copy-paste so there's no mistake; NERC too. At UKDS, suffix is based on internal ID; at ADS the suffix is sequential. Many institutions are using a combination of these approaches.
- A test account is possible with no obligation, but don't use the test DOIs publicly.
- Q: Is there a reason for optional metadata [i.e. beyond the minimum mandatory set]? A: We serve up your metadata online – all metadata goes into our service. It's very useful for harvesting, and there's a benefit of being able to talk about your data as widely as possible.
- Q: Cost? There's an annual fee which DataCite prefers to discuss directly with institutions for clarity. The contract is initially for three years then rolling annual.
- DataCite doesn't promote metadata to one particular place – DataCite pushes metadata out to wherever researchers are going.
- Institutions have prefixes. Possible to look up to whom the prefix belongs. DataCite could work with Crossref to see if they can do that yet, too.

Verena Weigert, Jisc. 'Jisc case studies on EPSRC compliance'
http://www.dcc.ac.uk/webfm_send/1930

- Verena, Monica Duke (DCC) and Jonathan Rans (DCC) have been working together to produce a report which they will publish at the end of March 2015, providing experiences from interviews with HEI staff who are working to meet the EPSRC research data policy requirements
- They hope this will be useful for other universities. The work has been undertaken with a view to sharing practice across the sector.
- In the report, the EPSRC requirements are grouped into three areas following DCC Cardio tool:
    - RDM policy, strategy, governance and sustainability
    - Support, RDM capability and skills

    - Tech infra and services required for storage, preservation and sharing

  Plus there will be challenges, quick wins and tips for other universities.
- Interviews have been completed with St Andrews, Leeds, UEL, Edinburgh.

Bill Worthington: 'RDM support services at the University of Hertfordshire'
http://www.dcc.ac.uk/webfm_send/1928

- Institution has around £15-20m research income per annum; keen to comply with EPSRC requirements and the general idea of reuse of research data.
- Our progress is a legacy of Jisc MRD programme work.
- We have a precarious advantage because there were 4-5 people during our MRD programme project, and then just me at the end of the programme (2013).
- We got a lot of useful outputs from the MRD programme: good strong policy, training materials, nascent systems, plus recognition that we needed to join this all up.
- A research data policy and a business case for sustainability of the research data activity is necessary, but it's more useful to build scalable solutions on a small scale and gather evidence of demand using them.
- Research data activity doesn't dilute the open access agenda; indeed, it can piggyback well on top of it.
- The RDM challenge brings together disparate elements of research support, both pre-award support and post-award, including DMP, finance, information services, storage, and other areas useful for good RDM practice.
- We have a new head of research and learning information services, who is RDM-aware.
- Subject information managers (used to be called subject librarians) does triage with all PIs for new projects.
- Triage signposts new project requirements and relevant service follow-up.
- Trying to make RDM business as usual with new researchers. Telling them they'll be publishing their data when they publish their papers. Data to go in subject archive if there is one, otherwise we'll look after it in the IR. Assumed to be open unless there's good reason. We inculcate this good practice across disciplines.
- We're doing things to mitigate the risks to data that can happen in universities: we're putting across strong messages in the following areas:
    - Networked storage:
        - Thousands of times more robust and safe than any local or portable device.
        - LAN: personal, departmental and research group levels available.
        - 128TB of tier 2 storage, cheaper and more flexible for research group use cases.
        - We have proved that commercial cloud storage can work, which probably means a sea change in autumn 2015 w Microsoft 365 Onedrive under Jisc framework agreement.
    - Document management system:
        - Enterprise document management system available for project work. For very high standard of data governance. Versioning, file level audited access, retention and disposal policies. Project-based folder template designed by UH researcher. Drag and drop via web GUI or mounted drive.
    - Open source Zendto for data transfer, deliver to collaborators:
        - Approved alternative to unregulated file sharing systems such as Dropbox, Google Drive, Yousendit, Mailbigfile, etc. Easier, cheaper and more secure transfer of sensitive material than existing practices such as those listed above, by email or by USB stick in envelope, particularly when sending sensitive information. Can take large files, and performs auto-disposal.
    - CRIS and IR:
        - Curation services by 2014: great for publications, inadequate for datasets. CRIS is Pure. Repository is DSpace. Last year they weren't working very well for data. Pure now supports datasets and metadata schema with 20 or so other agreed, and supports DataCite DOI minting. Repository is now working on relatively low-cost very long-term storage from Arkivum We're anticipating 10% of our research data needs to be kept in our system. Our OA repository is robust, elastic, low cost/TB/yr storage. We anticipate it will be in production in May 2015.

- Datasets@UHRA: a solution from 25-30K for the first 5TB for ten years. We can purchase more as demand and finance allow. For that you have a repository solution that will scale.
- 2 years ago – some slight progress on tiered storage model and use cases. All that progress came out of our Jisc MRD project.
- The journey benefited hugely from Jisc MRD and the infrastructure necessary for REF.
- Our journey has been inhibited by immature technology with components evolving at different rates. But now the technology's just about there to do the curation.
- EPSRC expectations are very useful despite having only 4 current EPSRC grant holders at our institution.
- We still have many issues to address, but the message is getting through. However, buy in from majority of researchers and money from the institution is required to sustain our progress.
- We will be ready and we can scale what we have according to demand.
- KA: Herts very pragmatic. Found a solution that will scale for the institution's needs, for the costs of 0.1% of their research budget.

Graham Blyth, Leeds: 'We can't even call it sensitive data'
Slides: http://www.dcc.ac.uk/webfm_send/1932

- In Leeds we don't use the word sensitive in our classification but there are issues and challenges around sensitive data for HEIs: classification, storage, access control, local contexts and how we share. Particularly re. data we can't store in our institutional repository.
- Classification schemes exist: DPA, military, government, industrial, consent forms, University. Each institution is working out how to apply these – should we do this together?
- Consent forms: researchers want to do the right thing. Vary in their approach. Don't want to put off potential participants so tend to be conservative and promise overly restrictive control on data. Need to work with ethics committees and share understanding and get it into the consent forms and DMP.
- Leeds University: data is unclassified, confidential or highly confidential. A lot of this has been done for institutional purposes, not with a research data focus.
- Assessing research data against classification schemes – mirrors DPA:
    - Leeds University: confidential means passport, home address, telephone number. Name plus whole address, etc.
    - Leeds University also gathers highly confidential data points: racial, ethnic, religious, sexuality, health, criminal record, academic progression.
- We should consider matching data assessment against repository classification and storage infrastructure elements.
- With a general-purpose repository most data can be open, and some can be made open via anonymisation or confidentialisation (from ANDS). Some can be controlled by authorisation of access, but some we probably can't even have it on our servers, let alone in a repository.
- Encryption helps – no longer data. OK with ISO27001. How do we give access in a controlled environment? This is something we're trying to do at Leeds.
- One path could be to encrypt, to put in a general repository. We can't manage the keys in the repository, but we have a new Integrated Research Campus and the Leeds Institute of Data Analytics, which can manage these things and might provide a place to allow controlled access.
- IRC never intend to be data repository. IRC does not hold archive. We take a copy over, give access, then delete from IRC when study completed. We don't hold it.
- Hearing about these offers today [i.e. from earlier presentations including from Jeremy Sharp] are also helpful.
- Longer term options: a national service? What would it look like?
- Summary:
    - understand classification schemes, consent forms.
    - Data assessment. Assessment of storage and repositories.
    - Access control for restricted, model for highly restricted.
    - What's the best ways to handle this? Let's delve into this community-wide.
- Q: If you don't allow confidential data on shared drive, how do you do health research? A: good point. We're actively engaged with our security people at the university. LIDA is going to have this data and have researchers working on it for that reason. However in many areas we're on the edge with the DPA.

- Comment: At a national level, there needs to be definition of classification of confidential.
- Show of hands – most in room interested in those summary questions.

Hardy Schwamm, University of Lancaster: 'Using Pure as data catalogue and (optionally) as data repository'
http://www.dcc.ac.uk/webfm_send/1931

- About 1/3 of audience using Pure
- Pure: 160 universities and HEI organisations in the UK.
- Of top 50 by research income, 43 have CRIS
    - Elsevier: Pure: 24 users
    - Symplectic: elements: ±20 users
    - Thomson Reuters: Converis: ±13 users
    - Eprints + extensions
    - Solutions developed in-house
- Pure doesn't interoperate with everything.
- Can use Pure as data catalogue.
    - Dataset now available as content type in latest pure release 4.20.
    - Dataset metadata template developed by pure user group
    - Dataset metadata and files can be displayed on pure research portal
- Data repository:
    - Files can be uploaded into Pure and are stored according to a mount point (server address)
    - Embargoes and access restrictions can be added
    - Pure has API and connectors so that data can by synced with other systems
- Two uses cases: Bristol and Lancaster data deposit proposed workflows. Bristol use as data catalogue; Lancaster use as data catalogue and repository.
- Pure user group (n=14) asked: using Pure as your data catalogue? 13 responses: yes, 1 response: no.
- Data catalogue live? Yes = 3, No = 11.
- Only one user has more than 50 datasets in their data catalogue.
- Using Pure as your repository? Yes = 8. No = 6.
- What do you think Pure does well with respect to RDM?
    - One stop shop / familiar for researchers
    - Good metadata schema
    - Supports compliance i.e. creates the link between the funding, the publications and the data.
- Biggest concerns
    - Commercial development
    - Not OS
    - Dependent on Elsevier priorities
    - Academics/researchers don't like it
    - User interface is muddled, inconsistent metadata fields
    - How can Pure interact with other university systems?
    - One size fits all metadata isn't suitable for all types of datasets
    - Lack of curation functionality
- How are you / will you be handling DP?
    - Don't know
    - Some level of preservation with EPrints, looking at links for long term dark archival storage points
- Lessons learned
    - We believe Pure can act as catalogue and data repository.
    - But there are limitations in interoperability of Pure with existing systems
    - And difficulties with its unpopularity with researchers
    - Further development is dependent on vendor
    - Pure user group represents community interests
    - Preservation is an issue that needs further investigation. Pure is not a preservation tool.
- Q: Joy: data validation stage on one of your slides. What sort of validation do you do? A: main validation will be linking the data with publications. Don't think we'll be validating the datasets – we don't have that expertise. But we'll check the fields are completed.

- Comment: The problems are not so much functionality of system but the usability of the system.

Wendy White, University of Southampton: A multi-pathway approach to RDM training?
http://www.dcc.ac.uk/webfm_send/1929

- We treat all training needs as an integrated whole.
- At Southampton, we had funding from the JISC MRD programme and we benefitted hugely from that.
- We aim to offer a one stop shop approach to advice.
- We involve later stage PhD students and ECRs in our RDM training, funded by the graduate centre, delivered within Library
- Like to use case studies and examples. Produced case studies as part of Jisc MRD project.
- Next one will be an interdisciplinary one with engineering and health. We are finding interdisciplinary approaches make useful examples / case studies.
- Our '101' starter course always has a waiting list. We need to look at our scalability as there is currently great demand for this training.
- There are some faculty-based courses – not disciplinary per se – but maybe not quite mapping how we'd like. Some faculty–related courses are busy and some not so busy.
- We view this as not a project, but rather day to day business.
- Aim: have overview of all elements; identify gaps. Identify any initiative that could be shared more widely. Start to map out pathways.
- Some of the options for training delivery include through pick and mix; intense and seasonal; emergency boost; integrated pathways.
- Reviewing elements of training: content, people, mode, time.
- We are developing or have developed curriculum modules about data analysis and ethics, as well as research methods courses.
- Specialist courses: software carpentry bootcamp, ADRC-E (Administrative Data Research Centre – England) courses on ethics preparation and analysis of data,
- Some of these potentially have wider applicability; we would encourage more sharing across disciplines. Let's make the most of the good things that are going on already.
- We are already promoting and re-evaluating existing course materials from the JISC MRD programme, to see if we can redeploy these.
- Map our practice to strategic development to institutional strategy.
- Making links between education and research activities. PGRs often straddle both e.g. research data and open access now including in postgraduate certificate of academic practice
- Experiential narratives informing podcasts, case studies
- International, sector and blended approaches to training targeted in order to develop and maximise expertise e.g. ADRC-E
- Some of the layers here are about how we hook into national and international offers.
- Q: point you made about having more data stories, please share them.

plus:
CLOSING COMMENTS: Kevin Ashley, DCC and Rachel Bruce, Jisc
What can Jisc / DCC do to help?

- Already available
    - Guidance regulatory environment: FOI, licensing, work with records managers, ethics committees, DMP online.
    - Data access including the research data discovery service, guidance on data licensing, example statements
    - Guidance based on the research concordat between universities and funders
    - RDDS: in Australia has helped standardisation data access statement.
    - Examples of policies and processes including institutional RDM policies.
    - DCC Institutional Engagement programme
    - Data storage framework agreements
    - CCEX tool – sharing costs – from 4C project
    - Benchmark figures
    - Drill-down on TRAC

    - Structured metadata: scientific metadata catalogue; DDS: going to gather evidence from which of those metadata fields help us with discovery. Guidance on data citation.
- WP in research at risk – N8 profile starting point based on recollect profile – see how that works as a recommendation that might be taken forward through via Casrai.
    - Evidence of researcher benefit of making data open and citable
    - We can provide training on-site.
    - Training – DCC, Jorum (which holds many of the Jisc MRD training outputs): library, IT, research office, researchers
    - Research at risk: usage stats; shared solutions; costing tools; work with RCUK on harmonised policy; intelligence gathering.
    - Case studies
    - Events like today
    - Institutional support
    - Cardio
    - Guidance

END