Directions for Research Data Management in UK Universities Meeting, Cambridge, November 2014

1. Introduction

1.1. Metadata

- Directions for Research Data Management in UK Universities Meeting, Cambridge, November 2014
  - https://docs.google.com/document/d/1fYgYkKX0NZlBE0KJtQu5iLcaCF3xk6c1695n4AMtmVQ/edit
  - http://bit.ly/D4RDM14
- 

This document is free of known copyright restrictions.
- Andy Turner's notes about a Jisc, RLUK, SCONUL, UCISA, ARMA and RUGIT meeting at the Moller Centre in Cambridge on the 6th and 7th of November 2014.
- Twitter hashtag:
  - #RDMRoadmap
    - https://twitter.com/hashtag/RDMRoadmap
- Main documentation
  - http://researchdata.jiscinvolve.org/wp/2014/12/04/directions-in-research-data-management/
  - http://www.jisc.ac.uk/events/directions-for-research-data-management-in-uk-universities-06-nov-2014

1.2. Contents

2. People

- Delegate list
  - http://jisc.msgfocus.com/files/amf_jisc/project_44/Directions_for_Research_Data_Management_in_UK_Universities_delegatelist.pdf
  - There may be additions and omissions
- Another list
  - Many of these were delegates
  - Lorraine Beard
    - http://www.library.manchester.ac.uk/aboutus/contact/staff-name-max-16-words-170680-en.htm
  - Graham Blyth
  - Rachel Bruce
    - http://www.jisc.ac.uk/staff/rachel-bruce
    - https://twitter.com/rachelbruce
  - Simon Dobson
    - http://www.simondobson.org/
    - https://twitter.com/simoninireland
  - Stephane Goldstein

        - http://www.researchinfonet.org/author/stephane-goldstein/
        - https://twitter.com/stephgold7
    - Barry Haynes
        - https://www.linkedin.com/in/barryrhaynes
    - Sarah Jones
        - http://www.dcc.ac.uk/about-us/dcc-staff-directory/sarah-jones
        - https://twitter.com/sjDCC
    - David Kernohan
        - http://www.jisc.ac.uk/staff/david-kernohan
        - https://twitter.com/dkernohan
    - Martin Lewis
        - https://twitter.com/MartinNHW
    - John MacColl
        - http://www.rluk.ac.uk/about-us/executive/
    - Stuart Macdonald
        - https://uk.linkedin.com/in/smacdon2
    - Mike Mertons
        - http://www.rluk.ac.uk/about-us/executive/
        - https://twitter.com/RLUK_Mike
    - Stephen Pinfield
        - http://www.sheffield.ac.uk/is/staff/pinfield
    - Rachel Proudfoot
        - https://www.linkedin.com/pub/rachel-proudfoot/41/6aa/aaa
    - Caroline Taylor
        - http://www2.le.ac.uk/library/images/staff/small/caroline-taylor/view
    - Peter Tinson
        - https://twitter.com/pat3460
    - Jo Wheatley
        - http://www.jisc.ac.uk/staff/jo-wheatley
    - James Wilson
        - https://www.linkedin.com/pub/james-a-j-wilson/2/b2a/5a6

3. Preparation

3.1. Draft Agenda

3.1.1. Day 1

- Introduction to the event
    - John MacColl
- The UK and Research Data Management: where are we?
    - Rachel Bruce
- What do disciplines do and what is the disciplinary research data management provision?

    - Sarah Jones, Stephane Goldstein
- Plenary discussion
- LUNCH
- The University of Edinburgh and the University of Manchester – how have they done it?
    - Stuart Macdonald, Lorraine Beard
- The data is the least of our problems
    - http://youtu.be/0j_GJr2gV8U
    - Simon Dobson
- Researcher needs – a researchers perspective
    - Andy Turner
- Reflections & questions
- BREAK
- Discussion groups
    - Five in parallel, two iterations
    - Focused on the key “pain point” areas:
        - 1) Policy implementation
        - 2) Skills and capability
        - 3) Infrastructure and interoperability
        - 4) Incentives - for researchers and for support stakeholders
        - 5) Business case and sustainability
    - What are the key issues that an RDM roadmap will need to address?
    - Facilitators:
        - David Kernohan
        - Mike Mertens
        - John MacColl
        - Peter Tinson
        - Rachel Bruce
- Group Facilitators (and rapporteur volunteers) will remain for a brief meeting.
- DINNER

3.1.2. Day 2

- Summary of day 1: what did we learn, what were the key areas of success and what needs addressing?
    - Mike Mertens, Caroline Taylor, SCONUL
- Where should collaboration happen?
    - Martin Lewis, Stephen Pinfield
- COFFEE
- Discussion groups:
    - A 5-year roadmap and vision
    - Facilitators
        - As yesterday.
- Plenary: actions and roadmap
    - John MacColl, Peter Tinson

- LUNCH and DEPART

3.2. Meeting location details

- Moller Centre
    - http://www.mollercentre.co.uk/mollercentre.php?lang=ENG
    - Map
        - https://www.google.com/maps/dir/Cambridge/52.214293,0.100004/@52.2039118,0.1015956,14z/data=!3m1!4b1!4m8!4m7!1m5!1m1!1s0x47d87083c513efe3:0xc3b0831300f8d3f1!2m2!1d0.137582!2d52.194571!1m0?hl=en-GB
- Accomodation
    - Felix Hotel
        - http://www.hotelfelix.co.uk/location
    - Map
        - https://www.google.com/maps/dir/Cambridge/Hotel+Felix,+Whitehouse+Lane,+Huntington+Rd,+Cambridge+CB3+0LX,+United+Kingdom/@52.2088567,0.0972513,14z/data=!3m1!4b1!4m13!4m12!1m5!1m1!1s0x47d87083c513efe3:0xc3b0831300f8d3f1!2m2!1d0.137582!2d52.194571!1m5!1m1!1s0x47d8772d45da4197:0x4c7ab40222436488!2m2!1d0.094413!2d52.224543?hl=en-GB

3.3. Presentation

3.3.1. Researcher needs – a researchers perspective

- Turner, A.G.D. (2014) Researcher needs – a researchers perspective. Presentation slides for the Jisc, RLUK, SCONUL, UCISA, ARMA and RUGIT Directions for Research Data Management in UK Universities Meeting, Cambridge, 2014-11-06, http://bit.ly/RDMRP14
- Outline
    - Researchers
    - Research
    - Research data
    - Researcher needs
    - Key issues

3.3.2. Personal experience/reflections

3.3.2.1. Key Issues

3.3.2.1.1. Ethics

- Ethical dilemmas and RDM
    - What do we do when we find useful data that perhaps should have been destroyed or never collected in the first place?
        - A topic of debate in psychology
- Governance

        - Are we allowed to use data collected for other purposes for a new purpose and what are the ethical controls on this?
            - The Information Commissioners Office guidelines have a principle that personal data should not be used for purposes other than what it was collected for.
                - http://ico.org.uk/for_organisations/data_protection/the_guide/information_standards/principle_5
- Freedom of Information (FOI) requests
- Demands for data access by authorities
    - How much confidentiality can researchers promise to research participants?
    - This is sometimes the reason why some personal research data is destroyed and the data is completely anonymised without record of who the research participants were, except that this knowledge can stay with the researchers that conducted the work.
- A key going forward is getting consent from research participants to share research data more broadly and to store personal data for re-identification under certain conditions and to enable research participants to be contacted for follow up studies
    - Finding research participants in the future could be a challenge, but with use of Personal Data Stores or via NHS registration or similar, this is becoming easier
    - In some definitions, personal data stops being personal when the person dies
        - Even if a person is dead, there can be close family that may be affected by disclosures

3.3.2.1.2. Research continuity

- Much more time can be lost finding things and figuring out how something worked, than would have been lost producing and linking metadata/documentation
    - Some researchers are over-confident that they will be able to do this
    - Some take shortcuts (risk and hope)
    - Some reason not to do this
        - Immediate survival is more important and there are more pressing matters
- Much of this depends on the people involved, their organising and the organisations, management and governance that should/could/might support them and their research.

3.3.2.1.3. Automation

- There is a balance between taking time to automate a complex workflow so that it is easy for things to be re-done and leaving in manual steps that have a greater demand on documentation.
- A working system that can be played with to see how it works is an order of magnitude easier to understand than one that is broken.
    - Documentation is also key

3.3.2.1.4. Digital preservation

- Use of open formats is very important

- Storing data in a database format is a danger
    - Storing data in a database might precipitate a need to store a copy of the database software and other aspects of the platform in order to be able to read that data in future
    - But sometimes the data is expensive to export out into a preservation friendly format
        - It takes effort to do this and automate loading into a database from these
- Some retiring researchers spend some time (often towards the end of their careers) developing the story of their career, in some cases making their research data and research findings more accessible.
    - This could be quite important.

3.3.2.1.5. Metadata

- Web Pages as metadata to
    - Most metadata can be open CCZero, some can't
    - Web Pages don't have to be open access
    - Key is to harness the power of Linked Data
- Adoption of metadata standards

3.3.2.1.5. Open source and ownership

- Not all research data can be made open data due to confidentiality issues
    - Much research data can though still be made available for researchers bound by the same ethical codes of conduct
        - It is a challenge though to support the interested public including people caring for or suffering from disease
    - Although it is important to know about what research has taken place, some research has to take place under the radar.
        - Do we know how much research is being done that the public are not allowed to know about yet?
            - Can FOI requests probe what is going on and find out more details?
            - What really needs to be embargoed to protect the research, those participating in it, those conducting it, supporting it and funding it?
- It is imperative to know about some research that has been conducted
    - For example clinical trials of treatments for diseases
    - Ben Goldacre in the All Trials campaign has argued the need for open access to clinical trial data
        - http://www.badscience.net/category/alltrials-campaign/
- These days academia is not where a lot of research is happening
    - There has always been research going on in more commercial enterprises and in government, but in recent years some big companies have led the field in research.
        - In some fields universities and academia struggle to keep up and even engage with the forefront of developments.

    - There can be advantages and disadvantages to opening up knowledge
        - Some companies need patents and the royalties they generate to keep going, so sharing all their trade secrets challenges them in survival and also pushes them to innovate more, but they can develop a competitive advantage by being ahead of the game...
- In science and in computational subjects it is becoming increasingly important to use only open source software tools and/or software that is standards certified to behave in a compliant and well documented way
    - This can be key to reproducing and validating results which may be essential in further research
    - With open source software, the internal workings can be scrutinised and minor changes can readily be made to produce new results for comparison.
        - This can help us to understand sensitivities and robustness

3.3.2.2. Personal experience

- This section focuses on past experience and is something of a reverse chronology.
- I like to think that the projects that I am now working on have much better RDM, but I know that we could still do much better in these, there are still risks that are taken, but so far I am getting away with it and we are thinking of also writing up our RDM experience as an example of practice to produce further research outputs that others might find useful.

3.3.2.2.1. MoSeS GENESIS NCeSS and NeISS

- Modelling and Simulation for e-Social Science (MoSeS) and Generative e-Social Science (GENESIS) were consecutively run nodes of The National Centre for e-Social Science (NCeSS) funded by the ESRC.
- The (National) e-Infrastructure for Social Simulation (Ne-ISS/NeISS) project began prior to the end of the ESRC funding for NCeSS and was funded by Jisc.
- NCeSS formed in around 2004 and experimented with the use of various portals and back end data storage infrastructure to support various e-Social Science research, some empirical, some methodological. NCeSS e-Infrastructure was further developed in the NCeSS e-Infrastructure project.
- An NCeSS Sakai Portal was developed with sophisticated role based access control security.
- As ESRC funding for NCeSS came to an end, the NCeSS Sakai Portal was supported via the NeISS project.
- Jisc (then JISC) funding for NeISS came to an end and the NCeSS Sakai Portal was unsupported
- A National Grid Service machine at manchester died and the NGS Oracle database which was used as a backend for the NCeSS Sakai Portal only persisted as database backup dumps.
    - I received a copy of the database dump file backups and with help got a Sakai portal instance up and running at Leeds so I could extract some data from some worksites

        - For security reasons, I was not permitted to make access to the portal generally available
- Some social simulation models that I had developed in GENESIS were ‘Grid Enabled’ as part of the NeISS project.
    - There were two types of model, one which tracked individuals day to day (a demographic simulation model), and another which tracked individuals second by second (a traffic simulation model).
    - The day to day demographic model focused on the processes of pregnancy, miscarriage, birth, death and migration, out of the two models this was the more developed and somewhat less challenging to ‘Grid enable’ (make it so that it could be run on and take advantage of distributed grid computers).
        - GridPP helped us to set up a Virtual Organisation (VO) for NCeSS to allow us to use the large scale computational infrastructure set up primarily for physics
        - Models were run for small test areas and results produced on ScotGrid site and compared with results produced on local computational infrastructure at Leeds.
            - The results were identical.
        - We scaled up to run larger simulations at ScotGrid, ones that took an order of magnitude longer to run on the local computational infrastructure at Leeds.
            - Again the results were the same.
        - We scaled up further to run 10 year simulations for Leeds
            - Migration was handled at census output areas level
        - We were scaling things up to run for all of England aiming to scale up further to run UK simulations when we ran out of time and funding.
            - This scaling up required the use of more LCG tools and more GridPP affiliated sites computational resources.
        - An application was made to ESRC to fund further development of this model to incorporate health status from a range of disease incidence and prevalence datasets (the logic being that health has a large effect on fertility and mortality probabilities), but this funding application was rejected.
            - Those of us researching this moved on to work on other things.
            - Many terabytes of data had been generated and was stored mainly on ScotGrid machines, but also at QMUL, Daresbury and Manchester.
    - There was nowhere to move the complete set of results we had generated.
        - For the largest simulation runs we extracted the metadata and some generalised model outputs which we hoped to use as the basis for some peer reviewed journal articles and which would allow us to reproduce the results in future (which we look forward to doing on more modest computational resources).

        - This work is now dormant and hasn't gone any further in the last 2 years
    - The work described above is only a small fraction of the work on GENESIS and NeISS.
    - A copy of all the project data for GENESIS and NeISS (except the full simulation model results) was stored on the NCeSS Sakai Portal
    - I don't have closure on this, I want to publish something further about the simulation models and the results, but time marches on...
    - For MoSeS I integrated some UK census data to develop individual level population data for social simulations
        - This data along with the Java programs used to generate it and metadata to help others recreate it from source was submitted to the UKDA
            - http://esds.ac.uk/DDI25/6763.xml
                - A minor success
            - ESRC and the UKDA pushed us to make these data available
                - There were concerns about getting the final payment from ESRC for the work and thoughts that the submission might have some bearing on the funding decision for GENESIS

3.3.2.2.1. Death of the CCG Webserver

- In 2001 the Centre for Computational Geography (CCG) experienced another catastrophe, the Web Server machine died
- Stan Openshaw (who lead the CCG up until a severe stroke in 1999) had realised that the advent and development of the World Wide Web changed everything.
    - In 1997 when I started working with Stan, we were encouraged to put as much information online about every CCG research project and to embrace the new way of publishing our work - online.
- Many CCG projects developed portals
    - The CCG Web Server became a cobbled together mess that worked
        - The configurations to support the portals were poorly documented and the development was ad hoc and unsustainable not least because portal frameworks were unstable, prone to having and developing security issues and need long term maintenance.
            - Overuse of too wide a range of web portal technology was probably in hindsight a mistake, but in playing with all this technology we learned...
    - There was no automation, and no easy way of setting these all back up.
- The CCG did have a backup strategy and we had complete backups of the CCG Web Server machine
    - With a new machine, in theory, we should have been able to get everything back up and running, but when the machine died there was not the time, resources or support to do this and it sadly seemed best to focus efforts elsewhere, particularly with ongoing projects.
    - I still spend some time every now and again remembering what was on the old CCG Web Server and making some of the portfolio available to the world again.

            - There is still a lot that I have not managed to surface and probably never will and I am one of the last around at the institution that really cares and remembers enough to do this…
                - The struggle against information entropy for some systems that hold it can only carry on for so long before it will not seem worth it anymore!

3.3.2.2.2. Stan Openshaw has a stroke

- https://en.wikipedia.org/wiki/Stan_Openshaw
- Stan, a hard working, well established researcher and revolutionary computational geographer suffered a severely disabling stroke in 1999 which led to his retirement.
    - Stan had founded and developed the CCG since 1992 when he took up post at the University of Leeds having previously been at The University of Newcastle.
    - Stan was a bit of a messy professor.
        - His room was a mess, his digital files were a mess. There were some CDs, lots of old disks and lots of teaching and research materials all over the place mixed up with personal professional correspondence.
    - Stan was doing lots of research at the time of the stroke and had a legacy of completed projects.
    - In the immediate aftermath of the stroke, the CCG pressed to continue the work that was on-going and hoped that Stan would recover and know where things were.
    - When Stan's recovery was only very partial, he was retired.
        - Stan's office was wanted and the CCG took on the task of clearing it out.
            - We went through the piles and started to reorganise and rationalise the collection
                - Deduplication saved a lot of space
                - Work was organised more chronologically and thematically and teaching, research and other materials were separated
            - Books were catalogued and kept as a collection for a while before being distributed to colleagues
            - The physical resource has been whittled down, but there are still several boxes in the archive containing details of old project and materials of potential historical interest
    - This was a tragedy on many levels.
    - In part, I am still doing what I am as I want to try and preserve Stan's research legacy and that of the CCG and to develop the subject that Stan pioneered and the CCG collectively developed.
        - Computational geography or GeoComputation is now mainstream and with with "Big Data" and generally, many are realising what the subject is and that computational geography is what they are dealing with.

3.3.2.2.3. Engaging with RDM experts at the University of Leeds

- In April 2013 I started observing the JISC RoadMap project

        - I was probably identified and asked to do this as a result of efforts to try to develop Linked (Open) Data at the University of Leeds
            - Leeds Linked Data: Case for support summary
                - https://docs.google.com/document/d/1QdQXRJUQqW2Ug7ewihYAv1lj4W3JbrT2dMrp9AANYq4/edit
            - Leeds Linked Data Business Case
                - https://docs.google.com/document/d/1iUMWsoPeP3DnYvj0tQILB7DzqHiduEE8JD6vwsS_Hj4/edit
- The work started in the JISC RoadMap project continued with some 'interim funding' and I contributed to this phase of developing RDM practice at the University of Leeds by keeping a track of what was going on, attending meetings and trying to plan and conduct a research data audit in the School of Geography to try to map research data assets.
- While a business case for developing an RDM Advisory Service was being developed I was seconded to work with Tim Banks and develop RDM practice for two faculties in the University of Leeds
    - This was hands on learning and involved developing data management plans and conducting research data risk assessments with other researchers.
    - This was a great time for networking and learning about how much geography goes on at the University of Leeds that does not involve its' School of Geography.
- I joined an Ethics Committee at the University of Leeds and learned what they were doing and how I could push through ethics for better RDM.
- I have done a lot of RDM reading and linked some details about most of this in a set of Google Docs documents. The master document is:
    - Research Data Management
        - https://docs.google.com/document/d/1EmZGB4cVyoK7WgL1sd2sgrmvChZPmYwTkjzzDRO91EU/edit

3.3.2.2.3. Empathy

- Open source
    - Developments are not without costs
        - You get what you paid for
        - Everyone can use what you paid for
        - You can use what everyone else paid for
- Thanks for all the efforts that are going on to improve RDM
    - Better RDM can enhance our research efforts and make them more efficient, it can also be key and enable research that otherwise would not be viable.

3.4. Reading/documentation

- http://recodeproject.eu/
- http://www.rluk.ac.uk/strategicactivity/strategic-strands/redefining-research-library-model/
- http://www.rluk.ac.uk/wp-content/uploads/2014/02/john-maccoll_collections-and-the-archive-layer3.pdf
- Background paper distributed in advance of the meeting by email as an OpenOffice Document "Cambridge background paper.odt"

    - Research data management – Synthesis of stakeholder interviews
        - Document prepared for “Directions for Research Data Management in UK Universities” event on 6th/7th November.
        - An adapted and abridged from a report synthesised from a series of eight interviews conducted by Sheridan Brown (Chygrove Consulting) on behalf of Jisc. These were detailed semi-structured interviews with key representative voices working with research data management in a variety of roles.
        - Essentially a snapshot of the current state of RDM in the UK university sector, highlighting where we are now and what remains to be done. It is provided as a starting point to discussion and debate; we would expect this workshop to enrich these initial findings and add additional perspectives. Following the workshop the outputs from that event will be melded with the information gleaned from the interview process and relevant background research to produce a concise report from which a roadmap may be fashioned.
        - Sections
            - Policy Development and Implementation
            - Skills and Capabilities
                - What core skills are required?
                - Skills for researchers
            - Infrastructure and Interoperability
                - What infrastructure is required for successful RDM implementation?
            - Incentives for researchers and support stakeholders
            - Business Case and Sustainability
                - Range of approaches
                - Costing Models
- Draft Concordat on Open Access Research Data
    - Distributed on paper at the event

Concordat On Open Access Research Data

Rick Rylance and Nick Wright

Introduction

Open Access to Research Data is at an early stage of development. The UK has existing strengths in this area but will need to develop these as times change. The intention [of the Concordat] is to establish sound principles which respect the needs of all parties. It is not the intention to mandate, codify or require specific activities, but to establish a horizon of expectation of good practice on data access.

It is recognised that there are a wide range of stakeholders – researchers, publishers, funders, research organisations, users (including public, business etc) and government.

These principles recognise that open can mean a range of things – for example, available and re-useable. Open access to research data carries implications for cost and there will need to be trade-offs that reflect value for money and use.

Draft Principles

1. Open access to research data is an enabler of high quality research, a facilitator of innovation and safeguards good research practice.
2. Open access to research data carries a significant cost, which all principles should respect.
3. There are sound reasons why the openness of research data may need to be restricted (e.g. verification, commercial confidentiality, patient privacy, excessive cost etc) but these must be justifiable.
4. Good data management is fundamental to all stages of the research process and should be established at the outset.
5. Curated data must be accessible, discoverable and useable.
6. The right of the creators of research data to reasonable first use is recognised.
7. Use of others' data should always conform to legal, ethical and regulatory frameworks including appropriate acknowledgement.
8. Data supporting publications should be accessible by the publication date and should be preserved in a citeable form.
9. Support for the development of appropriate data skills is recognised as an obligation for all stakeholders.
10. Regular reviews of progress towards open access to research data should be undertaken.

3.5. Liaison with colleagues

- I informed the University of Leeds RDM Advisory Service team about my invitation to this event
    - They provided feedback on my draft presentation slides
    - (And have covered my travel expenses)

4. Notes

4.1. Introduction to the event

- John MacColl
- A handpicked selection of some of the finest minds in the country from different types of groups engaging in RDM
    - Librarians
    - IT experts
    - Research administrators
    - Researchers
- Research libraries
- Engagement
    - Finding out what works
    - Developing practice based on what works
- Archives
- Digital humanities
- Data scientists
- Curating the products of scholarship
- We are aiming to prepare a Roadmap over the next two days

4.2. The UK and Research Data Management: where are we?

- Rachel Bruce
- Policy drivers and context
    - Definitions
        - EPSRC emphasise validation
        - H2020 definition considers the range of research data
    - Policy and interpretation vary
    - http://www.rcuk.ac.uk/research/datapolicy/
- Rocket Science (2011) Capacity builders Social Enterprise Programme
- Evaluation Regional Report
    - http://www.rocketsciencelab.co.uk/pdfs/SEP_regional_report.pdf
    - Science as a social enterprise report
- The UK is leading thinking about innovative RDM policy development in Europe
- Open data
    - Confidentiality
    - Intelligent openness

- UK Open Research Data Forum: Research Data Concordat
    - http://www.universitiesuk.ac.uk/highereducation/Documents/2012/TheConcordatToSupportResearchIntegrity.pdf
    - Single side A4 paper document distributed as draft
- Recognition that organisation at the UK level will take time
- Important things to push
    - Research data citation and metrics
- RDM Support Service Diagram
- Reflection on the different scales of this
    - International
    - National
    - Regional
    - Institutions
- DCC Curation Survey (from 7 months ago)
    - http://www.dcc.ac.uk/blog/rdm-2014-survey
    - Ethics confidentiality and FOI are tricky
    - Business planning and sustainability of RDM practise
    - Influencing higher managers was a big gap at the time
- Gaps
    - The DCC Curation Survey shows that there are lots of gaps and potential for collaboration between institutions
    - Research at Risk co-design challenge
        - Identified
        - Storage
            - Shared infrastructure for rapid deployment when needed
            - Arkivum
            - We need more sophisticated solutions
            - Replication
            - We might also need some different solutions for
        - Metadata
            - Range of views about how much of an issue this is
            - RIOXX profiles
            - This is both a solution and problem area
            - The problems are things like searching, retrieving, validating which are solved by metadata
            - Standards and minimal requirement guidance are needed
        - Preservation
        - Defining compliance
        - Moves towards automation
        - There are more questions than answers that came up in the consultation, we should spend some time reflecting on the questions and trying to find answers

        - The wheel is slightly broken, but we need to roll with it while we develop a better replacement that allows us to go further.
- Emerging solutions
    - Ease of use is key
    - DCC
    - Sherpa/Juliet
    - Janet shared data centre
- Time pressure to develop solutions
- Are we obsessed by compliance?
    - Is there then a risk that we miss some innovative solutions that might be there or that could be developed?
- Do we need to think outside the box?
- RDM policy and development is both a challenge and an opportunity

4.3. What do disciplines do and what is the disciplinary research data management provision?

- Sarah Jones, Stephane Goldstein
- VADS
    - http://www.vads.ac.uk/
- Kaptur
    - http://www.vads.ac.uk/kaptur/
- Kultivate
    - http://www.vads.ac.uk/kultur2group/projects/kultivate/
- Importance of personal websites for many researchers
- Data services
- Experiences of losing data is a driver for using (institutional) repositories
- Use of third party web services only is a risk
    - Terms of use changes
    - Service reliability and decommissioning
- Considering different subject areas
    - Digital humanities
        - More use of standards, XML and Text Encoding Initiative (TEI)
        - Example projects
        - Mapping Edinburgh's Social History
            - http://www.mesh.ed.ac.uk
    - Social Sciences
        - Greater awareness and acceptance of RDM
        - DDI
            - http://www.ddialliance.org/
        - Strong data centre infrastructure
    - Public Health
        - Ethics
        - Data Integration, linking and sharing

        - Longitudinal studies
        - Example project
            - Twenty-07: Public health study
                - http://2007study.sphsu.mrc.ac.uk/
    - Life sciences
        - More resistant to sharing
            - Fear of misuse
        - Less issues with resourcing
    - Genetics
        - Very rapid growth in data
            - Doubling every 6-8 months
        - Genbank
            - http://www.ncbi.nlm.nih.gov/genbank
        - ELIXIR
            - http://www.elixir-europe.org/
    - Neuroscience
        - Ethics
        - More data intensive and computational approach is being adopted
        - Lack of repositories
            - Much data stored at the research group level
        - Monash Open Microscopy Environment (OMERO)
            - http://www.openmicroscopy.org/
        - Sustainable disciplinary approach
    - Science and Engineering
        - RDM built in as standard
        - Established accelerated e-Research process of research
        - Commercial sensitivities
        - Mechanical engineering
            - Acceptance for archiving data
            - Still issues with preservations
            - Less sharing of data due to commercialisation
        - Crystallography
            - Much more open
            - Chemistry advocating open science approaches for a long time
            - Archives and national services well established
        - Astronomy
            - Less IPR issues, lots of sharing
            - More international sharing
            - Galaxy Zoo citizen science started here
            - Labtrove
                - http://www.labtrove.org/
                - Lab notebooks will become known as digital notebooks
- RLUK

- Research data typology
  - Helping librarians communicate with researchers with regards to RDM
    - Help through practical action (working together) classifying data
  - http://www.powtoon.com/show/fZDm1s0W6TI/research-data-typology-for-rluk-draft/
- Data categorisation
  - MANTRA
    - http://datalib.edina.ac.uk/mantra/
  - RDMRose
    - http://www.sheffield.ac.uk/is/research/projects/rdmrose
  - Re3
    - http://www.re3data.org/
    - Schema for describing and classifying repositories
      - http://www.re3data.org/2014/09/rfc-schema-version-2-2/
  - DCC Data Asset Framework (DAF)
    - http://www.dcc.ac.uk/resources/repository-audit-and-assessment/data-asset-framework
- Broad data types
  - How do researchers do it?
    - Data source
    - Provenance
    - Ready-ness of data for reuse
    - Purposes for accessing and reusing data
  - Media and volumes
    - File types
    - Volumes
  - How is data stored, managed and curated?
    - Use of metadata standards
    - Licensing and legal aspects
- An expandible resource
  - Crowdsourcing element
- Research data typology table v4 alternative view a pivot table Microsoft Excel document
  - Known that this should probably turn into Linked Data

4.4. Plenary discussion

- DMP plans and harmonisation
- Realisation that research data and RDM should be linked data and organised via a Resource Description Framework (RDF) approach
  - Semantic wiki anyone?
  - Co-production could be key
- The number of different types of data is growing exponentially
- There are file format registries

- Who is going to use this?
- What are we trying to achieve?
    - How common is this to different disciplines?
- The agile approach is a winner
    - We need to be aware of what is going on in different disciplines and not try too hard to shape this from the top
    - Engaging with practical help and support is perhaps the best way forwards
- How much is in common?
    - Perhaps 80%
        - There has been research on training materials that has found that there is on average something like 80% of the content is common
- Preservation of the software
    - Archiving this is not straightforwards
- Librarians having confidence to engage with researchers
    - It can be hard for librarians to deal with multiple disciplines

  Some research support staff lack confidence in engaging and reaching out to researchers
- Looking from the research process point of view is important to design the systems
- Do we consider all research outputs as research data?
    - I think so, but I was the only one to raise my hand.
    - Everyone agrees it is in terms of text mining
    - Concerns about regarding documents as data
        - Conflation will cause an issue in terms of all the other things being discussed in this space, perhaps in particular: open access.
- 'Trusted'/'recommended' repositories
    - http://www.datasealofapproval.org/en/
    - http://www.trusteddigitalrepository.eu/Site/Trusted%20Digital%20Repository.html
    - http://www.nature.com/sdata/data-policies/repositories
    - http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Guidance-for-researchers/WTX060360.htm
    - http://www.ed.ac.uk/schools-departments/information-services/research-support/data-management/data-sharing-guide

4.5. The University of Edinburgh and the University of Manchester - how have they done it?

- Stuart Macdonald, Lorraine Beard
- RDM Programme @ Edinburgh - an institutional approach
    - Edinburgh one of the first universities to adopt a RDM Policy
        - Aspirational May 2011
    - Governance
    - http://edin.ac/1eE3sav
    - DataShare
        - DSpace based multidisciplinary data repository

- Data archive
  - Vault
    - In development
    - Private stores
    - Discussions with Arkivum
    - Working on interface to incorporate user requirements
- Data asset register
  - Utilising PURE
  - Interoperation and coordination between all the service elements orchestrated around this.
- Customised instance of DMPOnline
  - Created their own templates for each of the funders
  - Tailored institutional Data Management Plan (DMP) boilerplate text
- DataStore
  - 0.5TB per researcher half of which can be pooled
  - Extra storage is available at £200 per TB per year
- RDM Support
  - 22 Schools with a representative library support person
- Awareness raising
  - Website, blog, wiki, face to face
- Training
  - MANTRA
    - Includes training kit for librarians
  - http://edin.ac/1kRMPv3
- RDM Programme
  - £1.2M
    - 75% infrastructure
    - 25% staffing
      - 7 people in total
      - Anticipating a growth in this
        - An average of 9.5FTE expected across Russell group universities for RDM in a few years time
- Current and future activity
  - DataSync
    - Secure dropbox like facility
    - About to be released
  - Developing performance indicators
    - Just enough rope for the RDM service deliverable managers to hang themselves!
  - DSpace Github
    - https://github.com/DSpace/
    - Lab notebook plugin
- Feedback

        - OwnCloud
            - http://owncloud.org/
        - DataVault
            - Discussion with Arkivum
            - Not sorted it out yet
        - All the Jisc funding developed material is creative commons attributed
            - These are all available
        - Researchers leaving data in DataStore rather than moving it to an archive
            - This is being left to the researchers who will need to move it to create space for new projects
        - There will be some funds available from Jisc for taking this forward
        - Git v Github
        - Steer from the management that researchers will not share software?!
        - New business case needed in 3 years
- Manchester
    - Mandatory policy for RDM for research projects with a Principal Investigator (PI)
    - RDM support service is a collaborative effort with key communities engaged
        - RDM service staff
            - 2 RDM core services
            - 12 in the Faculties
        - Developing metrics about service use to plan things going forward
        - Developed their own DMP planning tool in house using Sharepoint
        - Researchers can only submit a research bid if they have a RDM Plan reference number
    - Support
        - Web site
            - Well used
    - Storage
        - Each research project gets 20TB
    - Looking at Arkivum for archiving
    - Metadata management is a big issue
        - No data catalogue
        - Looking at using PURE
    - Service support challenges
        - Coordination across a large organisation
        - Fast changing landscape
        - Future proofing and scaling the service
        - Technical challenges
            - Range and scales of datasets
            - People still wanting to use a subject repository as well
    - Feedback
        - How much uptake is there on the 20TB and how is it costed?
            - 20TB funded through direct costs

        - Depositing in multiple places
            - How do you handle the submission?
            - How do you cope with the additional data added by users and repository managers and metrics?

4.6. The data is the least of our problems

- Simon Dobson
- http://youtu.be/0j_GJr2gV8U

4.7. Researcher needs - a researchers perspective

- Andy Turner
- http://bit.ly/RDMRP14
- I hope this conveys empathy and thanks from one researcher about all the efforts going into improving RDM
- Feedback
    - De-identifying/anonymising data: how hard is it and when is it appropriate?
        - Some data can be difficult and costly to anonymise
            - Audio and visual records can be particularly hard to anonymise
                - Does the data need distorting/perturbing to mask specific individuals identities?
        - We have to think carefully about whether de-identification/anonymisation is appropriate
            - Even if someone consents to going on the record we may still have an ethical obligation to protect their identity
            - Do we really have to promise to anonymise data in order to get consent from potential research participants?
            - If we cannot re-identify a participant that may be at risk of something and who we might want to contact again for this or other reasons, this can be a problem.
                - Sometimes it is better to retain a copy of all data prior to de-identification/anonymisation, push the data through a pseudonymisation process and make this more readily available...
            - Getting consent right is key and consent can be gained iteratively
                - Research participants need to be able to trust universities and their researchers, but there are higher powers and we cannot promise not to disclose any participants identity to all authorities if the personal data is kept
                    - This is why sometimes research destroys the personal data, though the identities of some research participants might still be remembered by the researchers.

        - Some anonymised data can be de-anonymised using other readily available data, or data that will become readily available
        - Some data can be aggregated into anonymised statistical forms and it can seem reasonable to release these data openly
            - e.g. Geographical statistics, such as population counts for even quite small spatio-temporal regions
        - Personal Data, Open Data, Data Sharing, Transparency, Pseudonymisation, De-identification, Confidentiality, Privacy, Big Data and Surveillance
            - https://docs.google.com/document/d/1FfoT-aZbcrNlPZp5SvGcH9jVluwbJX5M4iBDENzTmEY/edit
    - Research continuity is a good way to sell RDM to research managers/supervisors and institutions
        - Often there is a gap between projects and different researchers undertake the projects
            - Without appropriate metadata/documentation it can take significant effort to find data and reestablish the platform upon which a further study is to build
    - Sharing research data early should accelerate research, but some will view this as a risk in the current metrics/credit system where research data sharing/citation is not regarded as highly as citation of data analysis/insight findings published in peer reviewed journals
    - Giving credit where credit is due is challenging, but getting this right can fuel research
    - How to encourage researchers to share their data?
        - If a data set which is a research output is re-used in much further research, it may become more impactful than any analysis/insight gleaned from the individual/group that originally generated/curated that research data/output
            - We considered the example of using well marked up transcripts of interviews and the audio and metadata to develop tools for developing (more automated) transcription services.
        - Help researchers consider the possibilities if others also shared more (and earlier)
            - If the positives to outweigh the negatives it should start to happen...
    - The soft skills in dealing with researchers are important as there are different types of researcher and if you can understand their attitude and their current role and status you can begin to predict what they might be worried about in terms of data sharing and with regard RDM generally at different stages in the research lifecycle.

        - Retiring researchers that want to better map and reveal more details of their past projects are often doing a great job iterating over things, digitising and developing metadata to improve the curation of research
            - Not every researcher gets around to this!

4.8. Reflections & questions

- ...

4.9. Discussion groups

- Five in parallel, two iterations
- Focused on the key “pain point” areas:
    - 1) Policy implementation
        - With Rachel Bruce
        - We developed a map on paper of the policy landscape and key issues with regards policy
    - 2) Skills and capability
    - 3) Infrastructure and interoperability
        - Repositories
        - HPC
            - Active data and archive data
            - Speed of writing and retrieving from different stores
        - UKDA
            - Tiered model based on levels of data reuse
        - Active data versus published data
        - Scales of repositories
            - National, regional, local
        - Registries and catalogues
        - How to decide what to keep?
            - Making the decision to tidy up or just keep
        - Triggering data sharing and data destruction
            - Triggers such as the death of a person versus time based triggers
            - Again this could be demand driven rather than periodically checking a death register
        - Telling a story from your data
            - Some researchers will want all the data from all their projects to be available as part of their portfolio
                - This will tell a story of their career and is something more commonly done towards the end of a academics career
        - Why do some research funders not have repositories?
            - EPSRC deal with such a variety of data that there is concern that it could be a waste of resources or too much of a job to set up an EPSRC repository for all the research projects it funds
        - Does it matter where it is?

            - Perhaps not, so long as it is discoverable and can be accessed?
        - Data ingestion issues
        - Consortia Advancing Standards in Research Administration Information (CASRAI)
            - http://casrai.org/
        - Metadata
            - Credit for citation
            - National subscription for datacite
        - Publishers and reproducible research
            - PLOS ONE
                - http://www.plosone.org/
            - Nature have a science data journal publication
                - http://www.nature.com/sdata/about
            - Open research computation
                - http://www.scfbm.org/series/ORC
            - JASSS
                - http://jasss.soc.surrey.ac.uk/
            - Is there a publisher that requires as part of pre publication peer review that results can be reproduced from source when they are derived by computation?
                - http://science.okfn.org/2013/10/18/its-not-only-peer-reviewed-its-reproducible/
                - http://www.sciencemag.org/content/346/6210/679.full
                - Victoria Stodden
                    - http://blog.stodden.net/category/reproducible-research/
                    - http://web.stanford.edu/~vcs/
        - Research data
    - 4) Incentives - for researchers and for support stakeholders
    - 5) Business case and sustainability
- What are the key issues that an RDM roadmap will need to address?
- Facilitators:
    - David Kernohan
    - Mike Mertens
    - John MacColl
    - Peter Tinson
    - Rachel Bruce

4.10. Group Facilitators (and rapporteur volunteers) brief meeting.

- ...

5. References/Links

- Research Data Management

    - https://docs.google.com/document/d/1EmZGB4cVyoK7WgL1sd2sgrmvChZPmYwTkjzzDRO91EU/edit
- Knowledge Exchange (2014) Sowing the seed: Incentives and motivations for sharing research data, a researchers' perspective
    - http://www.knowledge-exchange.info/Files/Filer/downloads/Primary%20Research%20Data/Incentives/Incentives_for_Sharing.PDF
    - This study, commissioned by Knowledge Exchange, has gathered evidence, examples and opinions on current and future incentives for research data sharing from the researchers' point of view, in order to provide recommendations for policy and practice development on how best to incentivise data access and reuse.
- http://researchdata.jiscinvolve.org/wp/2014/12/04/directions-in-research-data-management/
- http://www.jisc.ac.uk/events/directions-for-research-data-management-in-uk-universities-06-nov-2014
- …