RENU Workshop: Open Access & Data Curation for Universities

Data Curation & Management Case Study – JISC projects and data services at the University of Hertfordshire

Oxford Brookes University, 17 September 2014

Bill Worthington, RDM projects Manager, University of Hertfordshire, w.j.worthington@herts.ac.uk
• JISC MRD Projects
• The data curation problem @UH
• A kind of strategy
• Get the message out
• Infrastructure
JISCMRD phase 2, 2011-2013

Two UH projects: infrastructure and short course in RDM

- JISCMRD programme designed to understand and build Research Data Management (RDM) infrastructure and capability

- Two projects, £300K JISC funding, matched by UH

- Cardio, DAF, work with ‘friendly’ research groups, much JISCMRD collaboration

- Outputs: systems, policy, guidance and training

- A legacy to build on, driver for an integrated research support service
The data curation problem

**Capacity** and Culture

JISC MRD Digital Asset Framework Audit: how **much data, how kept**, how organised, how shared?

- 12% return from research community, scaling the responses gives an estimate of 2PB (PB = million GB) of research data
- 10 x more than current central resources
- 80-90% in the hands of well resourced STEM research groups
- remaining 200 - 400 TB held by non technical researchers
- a lot of data held on workstations, laptops, local ad-hoc storage, insecure portable media, unregulated ‘free’ cloud services - and thus subject to considerable risk
- no sense of how much needs to be preserved

Capacity problem – say we have to curate 10% of ‘active’ research data for at least 10 years – we have no where to put 200TB and the systems where we might put it cost too much – **we need a low cost archive solution**
The data curation problem
Capacity and Culture

JISC MRD Digital Asset Framework Audit: how much data, how kept, how organised, how shared?

• archived on defunct media on the shelf, under the desk, in a cupboard
• little understanding of metadata
• collaboration only between trusted peers
• possessive competitive culture in many subject areas
• very little data deposit other than in disciplines that are already required to do so by their funders (eg NERC / British Atmospheric Data Centre)
• data protection often used to as an excuse not curate

Cultural problem – academics don’t know how or why they should publish data – we need to clarify policy, give clear advice, and appeal to the self interest of our researchers
A kind of strategy

Build (and threaten, cajole, or plead) and they will come…. 

- provide a data curation infrastructure
- don’t worry about the potential costs of success now, just make it work

and

- strengthen policy
- give practical assistance
- piggyback on existing practice
- exploit the open access agenda
- exploit any and all existing training and CPD channels to reach both senior and junior researchers
Policy
Strong statements with regard to open data

University of Hertfordshire’s Data Management Policy:

• ‘recognises the value of data as an institutional resource and considers that value to be increased through the widespread and appropriate use of data and by virtue of data quality’

• ‘considers the value of data to be diminished through misuse, misinterpretation or unnecessary access restrictions’.

Appendix III - Research Data:

• expects research data to be managed and shared in a robust and professional manner

• places the responsibility for proper research data management with the Principal Investigator (who becomes a Data Steward)

• all funded research projects to have a data management plan (DMP)

http://sitem.herts.ac.uk/secreg/upr/IM16.htm
Pre-award

- Research Grants Office assist with preparation of a DMP as part of the grant application process
- UH template for DCC DMPonline tool

Post-award

- Information Managers conduct an RDM triage with PI’s for all new projects
- Onward referral to technical teams
- Induction and refresher training on research systems
Training
Use every established channel

• All publishing researchers have to attend RIS training to get a seat at the system
• Post Graduate Researcher Development Programme
• Induction for new research staff
• CPD for early career researchers
• Research Institute workshops
• Open Access workshops
Cultural Change Message
Extend the ‘business as usual’ of publication to include data

Extract from Post Graduate Researcher Development Programme:

The selection of data, methods, algorithms, results, plots, and conclusions are in papers, published in journal….

…there is strong push toward open access for research papers….

…there is a corresponding requirement for open data to support those publications….
Consistently offer a default curation position
Selection, Placement, Access

Selection

• *that data which is required to support, reproduce or validate a publication*

• *(the EPSRC line in the sand)*

Placement

• in an appropriate national or international subject archive

• otherwise in the University of Hertfordshire Research Archive (UHRA)

• metadata held in current research information system (RIS) and available via Researcher’s public profile with other research outputs

Access

• assumed to be open

• ok in a restricted access repository with good reason, for example, for sensitive data
Existing Infrastructure @UH
Good for publications, inadequate for data

CRIS, Atira (Elsevier) PURE v4.n
- > 30,000 records
- 500+ researchers trained
- used to deliver REF2014
- portal component serves personal profiles with overview, activities, cv’s and research outputs
- poor dataset support

Repository, dSpace v1.8 (maintained by @Mire)
- > 12,500 research outputs
- populated via RIS
- delivers binaries of fulltexts and non-textual outputs
- machine harvestable dublin core metadata
- modest dataset support
- uses very expensive tier 1 data storage (share of MIS SAN provision)

RIS Portal (open) http://researchprofiles.herts.ac.uk/portal

UHRA (open) http://uhra.herts.ac.uk
New Infrastructure @UH
Soon with datasets!

CRIS, Atira (Elsevier) PURE v5.2.n
- Datasets distinct from other outputs in the data model
- Dataset metadata schema due early 2015, agreed by 20+ UK universities

Repository, dSpace v3.2 + Arkivum A-stor
- makes use of relatively low cost very long term storage
- datasets added directly to UHRA (not via RIS)
- information professionals will assist
Arkivum A-Stor is effectively infinitely elastic, and far more robust and cheaper than any single institution could build.

datasets@UHRA: a solution for £25-30K (for the first 5TB for 10 years). We can buy more as demand and finance allow.
In summary

• not yet established, but close enough to touch

• benefited hugely from JISC MRD, and the infrastructure necessary for REF

• hindered by immature technology, with components evolving at different rates

• new facilities ready in 2015

• still many issues to address

• needs buy in from majority of researchers and $$ from senior management
RDM resources @University of Hertfordshire

http://www.herts.ac.uk/rdm - RDM microsite

http://uhra.herts.ac.uk/handle/2299/13636 - JISCMRD final report

http://bit.ly/1pTPqWV - JISCMRD project blogs

http://go.herts.ac.uk/bill_worthington - other presentations on this theme