DCC Research Data Management readiness workshop

RDM support services at the University of Hertfordshire

13 February 2015, Paddington

Bill Worthington, University of Hertfordshire, w.j.worthington@herts.ac.uk
• The legacy of JISC MRD and a kind of strategy
• Research Support Service
• RDM systems
• Paddington 2013 reprised
• Ready to scale
JISCMRD phase 2, 2011-2013, two UH projects

A precarious but definite advantage

Legacy of JISCMRD @ UH:

- project team dispersed or departed
- an institution preoccupied by the REF
- RDM to be delivered as business as usual

- continued commitment of CIO, CTO, other influencers
- a good understanding of ourselves and the institutional task
- some outputs: policy, training, and nascent systems
- modest resources put aside for systems development

- recognition of the need to make the most of what we had; and for a more joined up approach to research support
- a new champion: new in post Head of Research and Learning Information Services
Meeting the RDM challenge
A practical strategy

Talk – and they will listen
Build – and they will come

Tactics:
• exploit all existing training, CPD and learned channels to reach both junior and senior researchers
• continue to strengthen policy
• exploit the open access agenda
• piggyback/hijack existing infrastructure and practice
• provide new scalable infrastructure
• offer all this plus ad-hoc practical assistance via a joined up Research Support Service

…. And don’t worry about the potential costs of success; demand is a good thing!
Meeting the RDM challenge
Research Support Service

The RDM challenge was the catalyst to bring together disparate elements research support

Pre-award support (finance and data management planning) + post award (information services, storage, good practice) + cpd & training

New in post Head of Research and Learning Information Services was RDM aware, and provided the glue by committing subject based information managers to the cause.

- Information managers carry out ‘research support triage’ with the Principal Investigators for all new projects.
- Triage signposts and identifies service requirements
- Service specialists follow up
RDM Training
Use every channel

• Research Information System training - all publishing researchers must attend 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

http://www.herts.ac.uk/rdm
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 these publications.
Consistently offer a default curation position

Selection, Placement, Access

Selection

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

Placement

• in an appropriate national or international subject archive
• otherwise in the University of Hertfordshire Research Archive (UHRA)
• metadata to be 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
• RDM systems
RDM systems
Networked Storage

Strong messages: network storage is 1000s time more robust than any local or portable device

Improved access: research students included, better VPN for off campus access, group working

Three flavours of LAN storage:

• User U:// private, not sharable, personal space for individuals (notional 5GB allocation, not hardwired)
• Departmental X:// school, dept, team folders, open by default, but access control to support research supervisor/student relationship
• Research R:// 128TB of tier 2 storage, cheaper, more flexibility for research group use cases

Also
• we have proved the point that commercial cloud storage can work
• probable a sea-change in Autumn 2015 with Microsoft365 OneDrive?
Enterprise document management system available for project work

- use when a very high standard of data governance is required
- versioning, file level audited access control, retention and disposal policies
- project based folder template designed by UH researchers
- Drag and drop via web GUI, or mounted drive
- Available to all PIs for group work
- may be defunct from Autumn 2015 with Microsoft365 Sharepoint on demand?
RDM systems
Opensource Zendto for data transfer, delivery to collaborators

• approved alternative to unregulated file sharing systems such as Dropbox, Gdrive, YouSendIt, and MailBigFile

• easier, cheaper and more secure transfer of sensitive material than existing practice (email or USB stick in an envelope)

• handles files too big for email

• record of transfer

• auto-disposal

http://zend.to
https://www.exchangefile.herts.ac.uk
RDM systems
Curation 2014: Great for publications, inadequate for datasets

Current Research Information System
Elsevier PURE v4.19
- > 30,000 records
- 500+ researchers trained
- very successful delivery of REF2014
- portal component serves personal profiles with overview, activities, cv’s and research outputs
- poor dataset support

Repository
dSpace v1.8 (maintained by @Mire consultants)
- > 13,500 research outputs
- mostly populated by the CRIS
- 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
Catalogue, Elsevier PURE v4.21/v5.2
- Datasets a distinct object type
- Metadata schema agreed by 20+ UK universities
- Supports minting of Datacite DOIs
- Some flexibility in attached storage

Repository, dSpace v3.2
- modified @Mire AssetStore module
- makes use of relatively low cost very long term storage (Arkivum100)
- Dataset binaries added directly to UHRA (not via RIS)
- archivist will assist
RDM systems - dSpace + Arkivum A-stor opensource repository + robust, elastic, low cost/TB/yr storage

- @MIRE AssetStore modified to use Arkivum API
- Dataset moves from dSpace to local A-Stor network appliance, then offsite to A-Stor tape cloud
- on request - first byte returns to local cache within 5 minutes
- dSpace GUI modified to support ‘request – wait (retrieving) – notify – deliver’ data journey
- ‘traditional’ outputs remain in modest sized local cache for immediate access
- Open/embargoed/restricted access states supported by dSpace 3.n

Excellent co-operation between Arkivum and @Mire

Arkivum A-Stor is effectively infinitely elastic, and far more robust and cost efficient than most single institutions could build

datasets@UHRA: a solution for £25-30K (for the first 5TB for 10 years). We can purchase more as demand and finance allow
Paddington 2013 reprised: progress on tiered storage model and use cases

**Tiered local storage**

- 2013, Monolithic SAN
- 2014, Tier 2 research group storage

**Cloud storage**

- 2014 Zendto
- 2015 OneDrive
- 2015 Arkivum A-stor

**Use cases**

- Desktop attached files
- Wide area access to personal files, sharing, repository
- Data at rest, backup, repository
- HPC, modeling
- Databases, CRIS

**Tiered local storage**

- Very fast
- Moderately fast
- Utility speed

**Cloud storage**

- Network limited
- Delayed on demand
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

RIS admin (private)

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

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

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
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
EPSRC readiness @ University of Hertfordshire

Closing comments

- EPSRC expectations very useful (despite having only 4 current grant holders)
- the journey benefited hugely from JISC MRD, and the infrastructure necessary for REF
- the journey was hindered by immature technology, with components evolving at different rates; a still many issues to address
- message getting through, however buy in from majority of researchers and $$ from the institution required
- but..... we believe we will be ready
- and.... we can scale what we have according to demand
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