DISC-UK DataShare project

DataShare & Data Audit Framework 2007-09: Lessons Learned

Robin Rice

University of Edinburgh, Scotland

Digital Curation

Practice, Promise and Prospects

Chapel Hill, NC USA April 1-3 2009

DISC-UK

Data Information Specialists Committee - UK

- A forum for data professionals working in UK Higher Education who specialise in supporting staff and students in the use of numeric and geo-spatial data.

- DISC-UK's aims are -
  - Foster understanding between data users and providers
  - Raise awareness of the value of data support in Universities
  - Share information and resources among local data support staff

- We are also members of **IASSIST**, an international organisation of professionals working in and with information technology and data services to support research and teaching in the social sciences.

DataShare project

DISC-UK has completed a JISC-funded repository enhancement project (March 07 - March 09) with the aim of exploring new pathways to assist academics wishing to share their data over the Internet.

With three institutions taking part – the Universities of Edinburgh, Oxford and Southampton – a range of institutional data repositories and related services have been established.

The project was led by the JISC-funded national data centre, EDINA, at the University of Edinburgh, which also runs the University's Data Library service.

Envisaged outcomes

- Exemplars of setting up institutional data repository services at each partner institution
- Enhancements to partners' IRs - with documentation and open source code for adapting DSpace, Fedora and EPrints repository software for handling datasets
- Toolkits, briefing papers and other outputs to inform UKHE repository community about data management and research support
- Technical watch on e-Research, VREs, Web 2.0 and related developments
- Papers, presentations and online dissemination of collected knowledge

Blogging and bookmarking

Actual live cloud tag is on project's *Collective Intelligence* web page based on social bookmarks collected on *Faves.*

Briefing Papers

- Gibbs, H. (2007). **DISC-UK DataShare: State-of-the-Art Review**
- Martinez, L. (2008). **The Data Documentation Initiative (DDI) and Institutional Repositories**
- Macdonald, S. (2008). **Data Visualisation Tools: Part 1 - Numeric Data in a Web 2.0 Environment; Part 2 - Spatial Data in a Web 2.0 Environment and Beyond**
- Green, A., et al (2009). **Guide to Data Requirements for Digital Repositories** (forthcoming)

Tech development: *Edinburgh DataShare*

- Upgrade to DSpace ver 1.51 with new theme aligned with University corporate style
- Registration streamlined using the University's single sign-on
- Embargo option - coded to restrict full data download with open metadata
- Open Data Commons license option (PDDL); else Rights field mandatory
- Date range enabled to allow Time Period (dc:coverage)
- Dynamically queries Geonames, a community generated database to find matching places & ensure consistency in metadata entry for Spatial Coverage field
- Extension to DSpace to record bitstream downloads in usage statistics
- Anti-virus checking upon upload
- Download All option (zip file of all item components)
- Citation field automatically generated based on specified metadata values.

Edinburgh DataShare Dublin Core-compliant metadata fields

Depositor (contributor)
Data Creator
Title
Alternative Title
Dataset Description (abstract)
Type
Subject Classification (JACS)
Subject Keywords
Funder (contributor)
Data Publisher
Spatial Coverage
Time Period (temporal coverage)
Language
Source
Dataset Description (TOC)
Relation (Is Version Of)
Supercedes
Relation (Is Referenced By)
Rights
Date Accessioned

Technical development: ORA

"A separate instance of Fedora has been set up which will act as a new repository running parallel to ORA - Oxford University Research Archive. It will hold research data and has been named *DataBank*. It is anticipated that DataBank will be a store for 'long tail' data ie data not held in other locations both within and external to Oxford University, and which does not comprise vast grid or similar datasets. In the first instance DataBank will not be directly accessible: access will be via digital objects held in ORA.

Content models for tabular data and for phonetic sound files in Fedora have been written."

Sally Rumsey, ORA Repository Manager

Tech development: ePrints Soton

"A prototype ePrints 3.1 repository has been created. This has enabled the deposit of research data, and associated metadata, under the new item type 'Dataset'. When the institutional repository, ePrints Soton, is upgraded to version 3.1 the data from the prototype will be transferred and become public. ePrints Soton will then be fully equipped to accept research datasets in addition to publications, patents, artefacts, exhibitions, musical compositions and performances." ***Harry Gibbs, Social Science Data Librarian, University of Southampton***

Also:

- Usability testing of data deposit interface with custom metadata fields
- Geonames lookup and autocomplete enabled by EPrints Services, with storage of longitude and latitude for display of location in Google Maps. *To be rolled out in next version of EPrints users.*

Partnerships in the Data & Research Lifecycle

Discovery and Planning

Data creation, collection, repurposing: *Partnerships between researchers & support services with subject expertise; informed by domain standards and guidelines relating to formats, metadata, version control, etc.*

Data Analysis

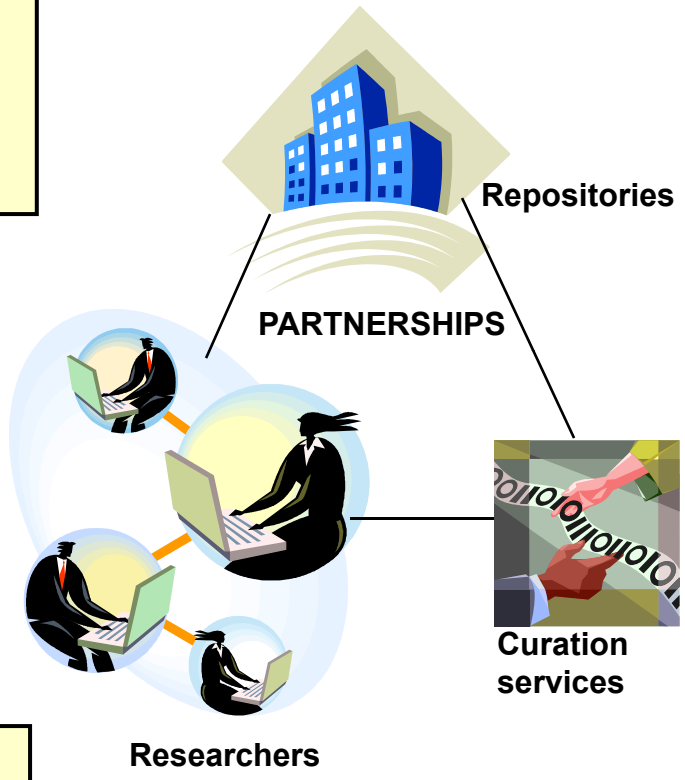
Data processing, management and curation: *Data are transformed, cleaned, derived as part of the research process; curators identify 'partnering moments' to capture content for documentation and description. Staging repositories offer curatorial workspaces.*

Publication and Sharing

Data sharing and distribution: *Repositories ingest and manage research outputs; offer federated searching, redundant storage, access controls; scholarly publications linked to data.*

Long term access

Data preservation, dissemination & long term stewardship: *Repositories and data archives provide preservation services such as format migration and media refreshment; dataset may survive a period of dis-interest before being re-discovered.*



Ann Green, Digital Lifecycle Computing

Enter Data Audit Framework

Recommendation to JISC:

"JISC should develop a Data Audit Framework to enable all universities and colleges to carry out an audit of departmental data collections, awareness, policies and practice for data curation and preservation."

Liz Lyon (2007). *Dealing with Data: Roles, Rights, Responsibilities and Relationships*

Data Audit Framework (DAF) Projects 2008

- JISC funded five six-month projects:
  - DAF Development (DAFD) Project, led by Seamus Ross (Director), Sarah Jones (Project Manager) HATII/DCC, University of Glasgow
  - Four pilot implementation projects:
    - King's College London
    - *University of Edinburgh*
    - University College London
    - Imperial College London
- Two more conducted by DataShare partners, the Universities of Oxford and Southampton, as added deliverables

DAF: Edinburgh conclusions

"…The audits were a good starting point and useful to identify the gaps and issues in managing data assets in the schools and units audited. Staff had numerous comments and suggestions for improvement of data management at different levels indicating an awareness of the issues, even where it has not been made a priority to address."

"While further awareness-raising is still important, staff require pragmatic assistance in the form of guidance on best practice, research unit or school procedures, College or University-wide infrastructure and policy, and identifiable forms of support for data curation in the form of expert support staff, web pages, and discipline-specific guidelines, as well as short, focused, training opportunities."

Cuna Ekmekcioglu and Robin Rice

DAF: Southampton conclusions

"… The data survey has demonstrated the need for a range of data support across the School, and the Library is preparing to respond. Initially, it is planned that support will take the form of a webpage containing links to relevant advice. There is also a role for more proactive data support that could be provided efficiently by the central Library service and this highlights the need to develop data management skills amongst Librarians." ***Harry Gibbs***

DAF: Oxford context

- The project *Scoping Digital Repository Services for Research Data Management* started in January 2008 as a cross-agency collaborative effort in Oxford. The project aimed to scope the requirements for digital repository services to manage and curate research data generated by Oxford researchers. The project contributed to the HEFCE funded UK Research Data Service feasibility study.
- As part of the requirements gathering exercise around 40 interviews with researchers took place and a consultation with service units in Oxford was conducted. The interviews with researchers helped us to learn more about their data practices and to capture their top requirements for services to support their data management.
- Overall, the Data Audit Framework proved to be an extremely valuable methodology to plan and execute a strategy to gather information about data management activities and data assets held within research centres in Oxford.

DAF: Oxford conclusions

The top requirements included:

- A sustainable infrastructure that allows publication and long-term preservation of research data for those disciplines not currently served by domain specific services;
- A secure and user-friendly solution that allows storage of large volumes of data and sharing of these with fine grained access control mechanisms;
- Advice on practical issues related to managing research data across the research life cycle.

Luis Martinez Uribe, Digital Repositories Research Co-ordinator

- See www.data-audit.eu
- DAF project reports available (findings)
- Appendices with questionnaires, interview schedules, etc
- Methodology document
- Online tool ready for others to conduct data audits

Lessons Learned Overall (1)

- Top-down drivers are important for overcoming barriers to data sharing (e.g. funders' requirements for data mgmt and sharing plans)
- Data management motivation is a better bottom-up driver for researchers than data sharing but is not sufficient to create culture change
- Institutional repositories can play a part in overall infrastructure for data sharing (see Data Sharing Continuum handout)
- Data librarians, data managers and data scientists can help bridge communication between repository managers & researchers (see Data Skills/Career study, Swan & Sheridan 2008)

Lessons Learned Overall (2)

- Institutions should consider developing research data policy, to clarify rights & responsibilities
- Institutions create a broad range of data in the course of research, not just rectangular datasets. So for *institutional* data repositories, the self-archiving model is probably the best for ensuring data quality. (Repository is a host, not a publisher. Only metadata is moderated.)
- IRs **can** improve impact of sharing data over the internet (permanent identifiers, citations, links with publications, discoverable metadata, long-term access and stewardship)
- Don't conduct institutional data audits unless you're prepared to open a can of data management worms!

Finally

- *And don't go it alone. Get buy-in from other institutional stake-holders (computing staff, librarians, department heads, principal investigators, records managers, archivists, research office staff). Collaborate. Have fun ☺*

www.disc-uk.org/datashare.html

R.Rice@ed.ac.uk