# Open Data and Institutional Repositories

Robin Rice
EDINA and Data Library, Information Services
University of Edinburgh, Scotland

CONFERÊNCIA SOBRE O ACESSO LIVRE AO CONHECIMENTO
University of Minho – Braga, Portugal 26 - 27 November, 2009

Overview

- Service background: EDINA & Data Library (a division of Information Services at UoE)
- Project background: DISC-UK DataShare
- Benefits to data deposit in an IR
- Policies for research data in IRs
- Towards a culture of (open) data sharing
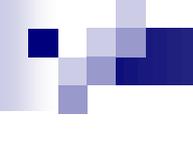
EDINA   JISC

EDINA is a JISC-designated national academic data centre based at the University of Edinburgh.

'Our mission is to 'enhance the productivity of research, learning and teaching' across all universities, research institutes and colleges in the UK.

'We do this by delivering first-rate online services … [for bibliographic, geographic and multimedia data] and by carrying out successful R&D projects.'

What is a data library?

A **data library** refers to both the content and the services that foster use of collections of numeric, audio-visual, textual or geospatial data sets for secondary use in research.

A data library is normally part of a larger institution (academic, corporate, scientific, medical, governmental, etc.) established to serve the data users of that organisation. The data library tends to house local data collections and provides access to them through various means.

Edinburgh Data Library services … distilled

- **Finding…**

"I need to analyse some data for a project, but all I can find are published papers with tables and graphs, not the original data source."

- **Accessing …**

"I've found the data I need, but I'm not sure how to gain access to it."

- **Using …**

"I've got the data I need, but I'm having problems analysing it in my chosen software."

- **Managing …**

"I have collected my own data and I'd like to document and preserve it and make it available to others."

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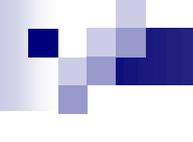
DISC-UK has completed a JISC-funded repository enhancement project (March 07 - March 09) with the aim of "exploring new pathways to assist academics wishing to share their data over the Internet" *(hint: repositories!)*

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.

Project Keywords

Capacity building
Best practice guidelines
Collaboration
Curation tools   Data archiving   Data Documentation Initiative (DDI)
Data formats   Data Mashups   Data professionals
Data publishing   Data sharing   Digital preservation   DSpace
EPrints.org   Funders' mandates
Edinburgh   e-Research   Fedora
Helping researchers   Institutional Repositories
Librarians   London School of Economics
Metadata standards
Managing research data
Open Access   Open Source Software
Open Data   Quantitative datasets
Orphaned datasets   Oxford
Research lifecycle   Skills & training
Secondary analysis
Southampton

"Live" cloud tag at http://www.disc-uk.org/collective.html
based on social bookmarks

Benefits to data deposit in an IR

- The repository submission process can guide the researcher through the process of data description to create a formal record of the dataset.
- IRs provide a suitable deposit environment where funders mandate that data must be made publicly available.
- Deposit in an IR provides researchers with reliable access to their own data for future use.
- Deposit of data in an IR, in addition to publications, provides a fuller record of an individual's research.
- Metadata for discovery and harvesting increases the exposure of an individual's research within the research community.

Benefits to data deposit in an IR

- Where an embargo facility is available, research can be deposited and stored until the researcher is ready for the data to be shared.
- Where links are made between source data and output publications, the research process will be further eased.
- Where the Institution aims to preserve access in the longer term, preservation issues become the responsibility of the institution rather than the individual.
- Time-stamps on submissions provide researchers with proof of the timing of their work, should this be disputed.

Gibbs, H. (2007). DISC-UK DataShare: State-of-the-Art Review

Policies for research data in IRs

- ■ Eligible depositors
- ■ Access and re-use
- ■ Data quality requirements
- ■ Metadata
- ■ Confidentiality and disclosure
- ■ Rights and ownership
- ■ Data file formats

*Green, A., Macdonald, S. and R. Rice, (2009). **Policy-making for Research Data in Repositories: A Guide***

Policies: eligible depositors

- Who is eligible to deposit data in the archive? (all staff? postgraduate students? others?)
- Will there be assisted deposit? (by whom?)
- Moderation by repository staff?
- What kind of data will be received?

… What kind of data?

- Scientific experiments
- Models and simulations
- Observations (surveys, censuses, voting records, field recordings)
- Derived data: resulting from processing or combining 'raw' or other data
- Multimedia: video recordings, images, sound

Policies: access and reuse

- Will access to the content in the repository be open to the public?
- Will registration be compulsory before downloading or accessing data?
- How will restricted access conditions be implemented? Will there be an embargo option?
- Will the archive or the depositor attach a license for re-use? (such as an open data license)
- Will a sample citation be provided for the dataset?

Policies: data quality requirements

■ What criteria do the data need to meet in terms of quality?

■ Is the coverage complete?

■ Have they been checked for validity?

See *DANS, (2008), Data Seal of Approval: Quality Guidelines for Digital Research Data in the Netherlands*.

Policies: metadata

- What metadata (descriptive, structural and administrative) should be supplied by the data creator?

- What happens if this required metadata is not supplied? What metadata will the archive supply?

- Are the data well-documented for secondary use? (Are variables and values labeled? Are there supplemental materials such as a codebook, protocol, technical report?)

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

Policies: confidentiality and disclosure

- What requirements must data creators meet regarding confidential data? (Is this explicit in the depositor agreement?)
- Does data, for example, identify people?
- Will the archive anonymise data?
- If data is not fully anonymised, have the subjects given informed consent for releasing the data?

Policies: rights and ownership

- What rights relating to the data are retained by the creator?
- What rights are transferred to the archive?
- Can the archive change the data at all (for example, during processing data for preservation)?
- Does the creator certify that the data do not infringe the copyright of others, or certify that permission from the rights-owner has been received?

Policies: data file formats

- Which formats will be accepted for deposit?
- Which formats are preferred?
- Will the formats be normalised by the archive?
- Will compression formats (for example, zipped files) be accepted?
- Will the archive retain the original bit stream as well as the normalised files?
- Is there a limit to the size and number of files the archive will accept?

Towards a culture of data sharing

- Top-down drivers are important for overcoming barriers to data sharing, as they are for open access publishing.
  - For example, funders' requirements for data management and sharing plans
- Efforts by institutions to help researchers improve their research data management practice can help create culture change towards sharing.
- Institutional repositories can play a part in overall infrastructure for data sharing (see Data Sharing Continuum).
- Data librarians, data managers and data scientists can help bridge communication between repository managers & researchers (see Data Skills/Career study, Swan & Brown, 2008).
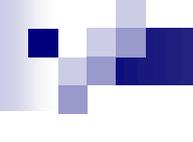
… A. Swan, S. Brown, 2008

The report calls for a 'repositioning' of the role of the library in data-intensive research. The authors of the report Alma Swan and Sheridan Brown write: 'We see three main potential roles for the library...Increasing data-awareness amongst researchers; providing archiving and data preservation services through institutional repositories; and developing a new professional strand of practice in the form of data librarianship.'

Thank you for listening ☺

Edinburgh DataShare repository:
http://datashare.edina.ac.uk/dspace

DISC-UK DataShare project:
http://www.disc-uk.org/datashare.html

Robin Rice, R.Rice@ed.ac.uk