DISC-UK DataShare

Robin Rice, Edina, University of Edinburgh

Introduction

DISC-UK DataShare (http://www.disc-uk.org/datashare.html) is a JISC-funded collaborative project led by EDINA and Edinburgh University Data Library, with the Universities of Oxford and Southampton as partners, and the London School of Economics as an associate partner. Our purpose is to develop models for the deposit of social science datasets in institutional repositories (IRs). The name of DataShare was chosen to be more or less self-evident; DISC-UK however, stands for Data Information Specialists' Committee-UK. DISC-UK members (a group of information professionals whose jobs are geared towards academic data support, primarily for the social sciences) believe that IRs may be developed to rescue some of this 'orphaned data' and make it available for future research, to the benefit of research communities and wider society.

Diagram of partner institution's experience in repository development projects and related services, including data libraries.

Each of the partner institutions has a strong commitment to enhancing their IRs, and DISC-UK members are working in partnership with IR managers at each of our institutions to develop solutions that work in each of the three main repository platforms, EPrints, DSpace and Fedora.

Research data: the case for curation

Liz Lyon observed in the JISC-commissioned Dealing with data report that, while many institutions have developed IRs over the last few years to store and disseminate their published research outputs, "…there is currently no equivalent drive to manage primary data in a co-ordinated manner" (p.45). Although policies and practices currently operate to gather, store and preserve data, chiefly in national, subject-based data centres, much data remains unarchived and is at serious risk of being lost.

In its Stewardship of digital research data report, the Research Information Network (RIN) examined the responsibilities of research institutions, funders, data managers, learned societies and publishers in turn. For example, research councils may choose to fund a domain data archive, as ESRC does for the UK Data Archive at Essex, or they may require grant applicants to include a data sharing plan, as the MRC has been doing since 2007.

DISC-UK recognises that the institutional role is just one part of the picture, and that libraries play just part of the institutional role, along with computing services and central research offices for example, but also recognises that IRs can play a role in improving the quality of data curation for at least some datasets, such as those that might be coupled with published research papers. As Prof. Peter Buneman, Research Director of the Digital Curation Centre is fond of saying, "If you want to preserve your research data, publish it!"

The Data Sharing Continuum graph below, developed as part of the project, shows how most datasets languish with minimum levels of management applied to them, while only a select few are given the highest levels of curation and prepared for publication. The ideal is to move everything up at least one level where possible.

From Open Access to Open Data

'Open Data' as a concept is quickly gathering momentum as the latest term in the 'open' trilogy along with Open Source and Open Access. It indicates a recognition that there is a rising level of expectation among users for complete access to an intellectual work, not only the final published post-print, but the body of evidence drawn on to create that final output. This is compatible with the scientific method of allowing replication of results by others, and the rich tradition of secondary analysis in the social sciences and other population-based research domains. It is also in line with recent initiatives to open up publicly-funded research data to public availability (for example, OECD, 2007).

Dr. Peter Murray-Rust, a chemist at Cambridge whose methods include mining chemical data from research papers has explained, "Where the Open Access movement is concerned only with ensuring that scholarly papers are human readable, the Open Data movement requires that they are also machine readable." The Open Knowledge Foundation states simply that "A piece of knowledge is open if you are free to use, reuse, and redistribute it."

In the same way that Creative Commons has helped overcome copyright barriers by providing authors with easy templates for licensing their work to readers, Science Commons has developed an Open Data protocol to facilitate machine-processing of data, to allow analysis, data manipulation and integration, and re-distribution in new forms, such as web 'mashups' in the Web 2.0 style. The principles behind the protocol are:

1. The protocol must promote legal predictability and certainty.

2. The protocol must be easy to use and understand.
3. The protocol must impose the lowest possible transaction costs on users.

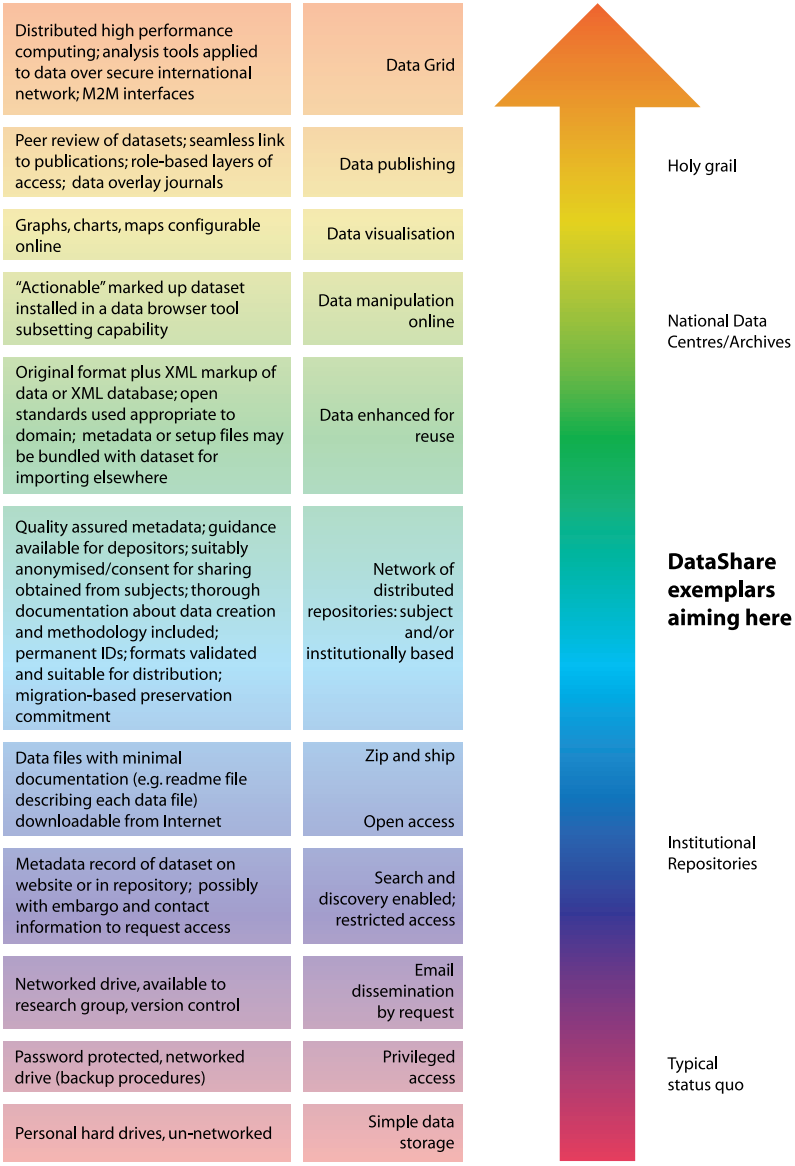
The protocol was followed in the Open Data Commons Public Domain Dedication & Licence (PDDL) written by Law researchers at the University of Edinburgh, which aims to bridge international differences in law to 'converge on the public domain' by waiving all rights based on intellectual property and to avoid 'attribution stacking' which produces a high burden on the user of data from many different sources. It also addresses the "sui generis" database right which applies in European jurisdictions.

There are of course reasons data are not always provided in a completely open way. Aside from commercial value, these may include confidentiality concerns, copyright questions, and sheer complexity. These concerns have been documented by the GRADE and STORE projects in the JISC Digital Repositories Programme, and summarised in DISC-UK's State of the Art Review on data sharing (Gibbs). Many scholars are more comfortable with an informal method of sharing, so they can assess the use to which the data will be put and decide whether to give the requestor access on a case by case basis. Allowing different levels of access based on data depositors' preferences is something the project partners are looking into with the LSE-based FAR project; for example, using the UK Access Federation method of authentication and authorisation control as a way of identifying members of a virtual organization who have been designated as having the right to access a particular dataset.

Assisting researchers early in the research process by providing guidelines for gaining appropriate consent from subjects, and anti-disclosure techniques such as anonymisation is also within scope of the project. Though we are not methodologists ourselves, there is solid information from authoritative sources that can be compiled and presented to researchers pro-actively to help them make informed decisions.

Meddling with metadata

Application of appropriate metadata is an important area of exploration for the project. Datasets are not different from other digital materials in that they need to be described, not just for discovery but also for preservation and re-use. The GRADE project found that for geo-spatial datasets, Dublin Core metadata (with geo-spatial enhancements such as a bounding box for the


Distributed high performance computing; analysis tools applied to data over secure international network; M2M interfaces
Data Grid
Peer review of datasets; seamless link to publications; role-based layers of access; data overlay journals
Data publishing
Graphs, charts, maps configurable online
Data visualisation
"Actionable" marked up dataset installed in a data browser tool subsetting capability
Data manipulation online
Original format plus XML markup of data or XML database; open standards used appropriate to domain; metadata or setup files may be bundled with dataset for importing elsewhere
Data enhanced for reuse
Quality assured metadata; guidance available for depositors; suitably anonymised/consent for sharing obtained from subjects; thorough documentation about data creation and methodology included; permanent IDs; formats validated and suitable for distribution; migration-based preservation commitment
Network of distributed repositories: subject and/or institutionally based
Data files with minimal documentation (e.g. readme file describing each data file) downloadable from Internet
Zip and ship
Open access
Metadata record of dataset on website or in repository; possibly with embargo and contact information to request access
Search and discovery enabled; restricted access
Networked drive, available to research group, version control
Email dissemination by request
Password protected, networked drive (backup procedures)
Privileged access
Personal hard drives, un-networked
Simple data storage
Holy grail
National Data Centres/Archives
DataShare exemplars aiming here
Institutional Repositories
Typical status quo

'coverage' property) was sufficient for discovery within a DSpace repository, though more in-depth metadata or documentation was required for re-use after downloading (Seymour). The project partners are examining other metadata schemas such as the Data Documentation Initiative (DDI) versions 2 and 3, used primarily by social science data archives (Martinez). Crosswalks from domain specialist schemas such as the DDI to qualified Dublin Core are important for interpreting the best use of DC metadata for description of research datasets at the study level (as opposed to the variable level which is largely out of scope for this project).

Building Capacity

One of the key aims of the project is to help to build institutional capacity for support of research data curation. Project staff are tracking developments nationally and abroad to inform librarians and others who share our interests on our website. Our collective intelligence section links to our blog, a live tag cloud of social bookmarks that amounts to a dynamic bibliography, and incoming newsfeeds from key sources. See http://www.disc-uk.org/collective.html.

We are also collaborating with others, such as staff at the UK Data Archive and the Digital Curation Centre, who share an interest in training and professional development for data curation.

A number of new data-related projects have been funded and JISC is organising meetings among them to ensure consistency with overlapping activities and aims. In addition to DataShare, these include:

- UK Research Data Service (UKRDS) funded by HEFCE
- Data Audit Framework Development Project (DAFD)
- Data Audit Framework (DAF) Implementation Projects
- Data Skills/Career Study
- DCC Digital Curation Summer School (DCSS:) *to take place in October 2008*.
- Preservation Costs Study

These are some of the important challenges faced by the project.

References

GIBBS, H. (2007) *DISC-UK DataShare: State-of-the-Art Review.* DISC-UK. http://www.disc-uk.org/docs/state-of-the-art-review.pdf

HATCHER, J and C WAELDE (2008). Open Data Commons Public Domain Dedication and Licence. Open Data Commons.
http://www.opendatacommons.org/odc-public-domain-dedication-and-licence/

LYON, L. (2007) *Dealing with data: roles, responsibilities and relationships*. UKOLN.
http://www.jisc.ac.uk/media/documents/programmes/digital_repositories/dealing_with_data_report-final.pdf

OECD (2007) OECD *Principles and guidelines for access to research data from public funding.* Paris: OECD. http://www.oecd.org/dataoecd/9/61/38500813.pdf

POYNDER, R. (2008). "The Open Access Interviews: Peter Murray-Rust." *Open and Shut* [Weblog]. http://poynder.blogspot.com/2008/01/open-access-interviews-peter-murray.html
RIN (2008) Stewardship of digital research data: *a framework of principles and guidelines*.
http://www.rin.ac.uk/files/Research%20Data%20Principles%20and%20Guidelines%20full%20version%20-%20final.pdf

SCIENCE COMMONS (2008). *Protocol for implementing open access data.* [Website] http://sciencecommons.org/projects/publishing/open-access-data-protocol/