DISC-UK DataShare Project: Building Exemplars for Institutional Data Repositories in the UK

*Robin Rice**

Introduction

DISC-UK (Data Information Specialists Committee - United Kingdom) is a forum for data professionals working in UK Higher Education who specialise in supporting their institution's staff and students in the use of data for analysis (primarily statistical and geo-spatial). This partnership, led by EDINA, is carrying out the DISC-UK DataShare project (March 2007 - March 2009)[1] that aims to explore 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 – plus the London School of Economics as an associate partner, a range of exemplars will emerge from the establishment of institutional data repositories and related services. It is part of a wider programme to develop institutional repositories funded by the UK's Joint Information Systems Committee (JISC).

This project brings together the distinct communities of data support staff in universities and institutional repository managers in order to bridge gaps and exploit the expertise of both to advance the current provision of repository services for accommodating datasets. The project's overall aim is to contribute to new models, workflows and tools for academic data sharing within a complex and dynamic information environment which includes increased emphasis on stewardship of institutional knowledge assets of all types; new technologies to enhance e-Research; new research council policies and mandates; and the growth of the Open Access / Open Data movement.

This article will summarise the work of DataShare in the following areas: defining the institutional data repository in the broader landscape and within the 'data sharing continuum'; investigating deposit of research data in institutional repositories including metadata and policy development; understanding and improving data management practice through partnering with academic departments in the use of the Data Audit Framework; and licensing issues including 'open data'.

Data and Institutional Repositories

According to the JISC-commissioned Digital Repositories Review (Heery and Anderson, 2005), a repository is differentiated from other digital collections by the following characteristics:

- content is deposited in a repository, whether by the content creator, owner or third party
- the repository architecture manages content as well as metadata
- the repository offers a minimum set of basic services e.g. put, get, search, access control
- the repository must be sustainable and trusted, well-supported and well-managed.

Institutional repositories are those that are run by institutions, such as universities, for various purposes including showcasing their intellectual assets, widening access to their published outputs, and managing their information assets over time. These differ from subject-specific repositories, such as Arxiv (for Physics papers) or RePEc (Research Papers in Economics).

The project – along with others funded simultaneously – will help to realise the vision of the Digital Repositories Review of a "coherent aggregation of content from a network of institutional repositories", and more particularly of the Digital Repositories Roadmap, e.g. the milestone under Data: "Institutions need to invest in research data repositories" (Heery and Powell, 2006).

There are of course some notable centralised data archives and centres serving particular disciplines in the UK, such as the UK Data Archive/Economic and Social Data Service (UKDA/ESDS) for the social sciences and the Natural Environment Research Council (NERC) Data Centres for natural and environmental sciences. Other disciplines have created vast online databases on the Internet or over e-Research grid networks, which is the logical place for 'publishing' data outputs in those domains. (Digital Archiving Consultancy et al, 2005; Swan and Brown, 2008a.) This project does not aim to challenge these nationally funded organisations that have set internationally recognised high standards in data archiving, management and curation, nor the model of domain-specific data archives/centres. It does, however, aim to explore the role of filling in the gaps left open by the paucity of coverage

of dedicated data archives, and in doing so, gain leverage from being able to work closely and directly with potential depositors at one's own institution. Indeed, the lifecycle approach to data sharing encourages intervention at the earliest stages of a research project to ensure adequate consent, documentation etc., are achieved for the data to be usable by others (Humphrey, 2000).

One of the first tasks of the DataShare project was to learn what repository managers had done in earlier projects and to what extent institutional repositories (IRs) in the UK were already dealing with data. A DISC-UK member therefore conducted a thorough State of the Art Review, which included depositors' motivation and barriers for depositing data in an IR (Gibbs, 2007).

The Data Sharing Continuum

Institutional data repositories are only one possible response to data publishing requirements of creators and funders, and they have limitations, such as lack of ability to visualise or manipulate the datasets online (using DSpace, Fedora, or EPrints software as it exists at present). The dataset, e.g. data files plus documentation, must be downloaded from the repository and analysed on a desktop computer using requisite software. This marks the "zip and ship" level of data sharing that IRs are well-suited to host. By adding value in terms of interacting with users to enhance metadata, documentation, and to reformat data into suitable sharing and preservation formats, we see our institutional data repository services as sitting comfortably in the middle of the Data Sharing Continuum shown below. The increased level of human effort required to curate data at the highest levels may be reserved for the most important, special, or highly-used datasets, as is done at national data archives. On the other hand, by raising awareness, providing local services, and offering a repository with a simple deposit interface, there is scope for the numerous datasets languishing on portable drives or with minimum bitstream backup only, to be moved up a notch or two on the scale, and therefore not lost to potential new uses and to the scholarly record (see figure 1 page 24).

In some cases it is the data creators themselves who wish to re-use the data later on, after they have moved onto other research projects. If they have documented and deposited their data for sharing purposes, they will not have the experience of many researchers of not being able to find, read, or interpret their data at a later time.

The project is looking at assisting researchers not only with depositing their data in an institutional repository or data archive, but also with using Web 2.0 tools to "mashup" their data for online visualisation through numeric-based applications such as Swivel and geo-spatial tools such as OpenStreetMap. There are of course advantages and disadvantages of using these type of "cloud" computing applications to publish academic data, which is covered in two briefing papers produced by the project (Macdonald, 2008a & 2008b).

Metadata and policy development

Application of appropriate metadata is an important area of development for the project. Datasets are not different from other digital materials in that they need to be described for discovery and also for preservation and re-use. The GRADE project found that for geospatial datasets, Dublin Core metadata (with enhancements such as drawing a bounding box to enter geospatial coverage) give sufficient context for discovery within a DSpace repository, though more in-depth metadata or documentation is required for re-use after downloading (Seymour, 2007).

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, 2008). Crosswalks from the DDI to qualified Dublin Core are important for describing research datasets at the study level (as opposed to the variable level which is largely out of scope for this project).

DataShare is benefiting from work of the Dryad project -- a repository for evolutionary biology[2] (Carrier, et al, 2007) and GAP[3] (Geospatial Application Profile) in defining interoperable Dublin Core qualified metadata elements and their application to datasets for each partner repository. The solution devised at Edinburgh for DSpace makes use of just twenty fields from Dublin Core (simple) and dcterms (qualified) and attempts to follow Dublin Core Metadata Initiative (DCMI) recommendations in applying them to datasets (Rice, 2008). One innovation introduced is to use a look-up to the open utility GeoNames for ensuring consistency in entry of placenames for geographic coverage.

DataShare also is developing a briefing paper[4] to provide a range of requirements that repositories can consider as they plan to add research datasets to their digital collections. The briefing paper discusses the scope of a data repository, its content policies for types of files and data sets held, its metadata policy for descriptive information about items in the repository: its submission policy, and its quality, copyright, and preservation policies. Background material was gathered from the online OpenDOAR Policies Tool[5] maintained by SHERPA at the University of Nottingham, the OAIS Information model and the TRAC checklist, and other sources.

Data management partnerships

The *Stewardship of Digital Research Data* report, (Research Information Network, 2008) 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 the Economic and Social Research Council

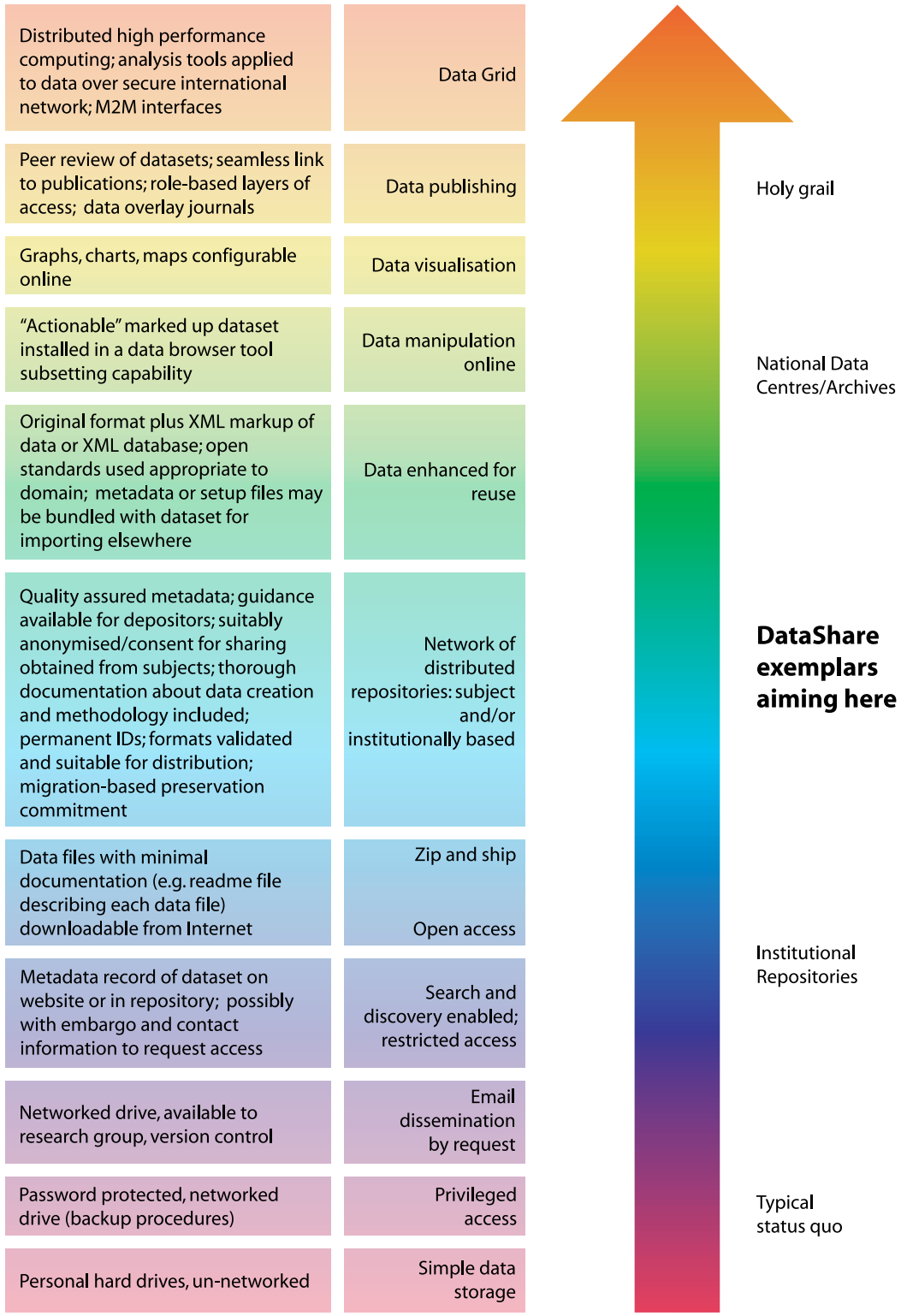
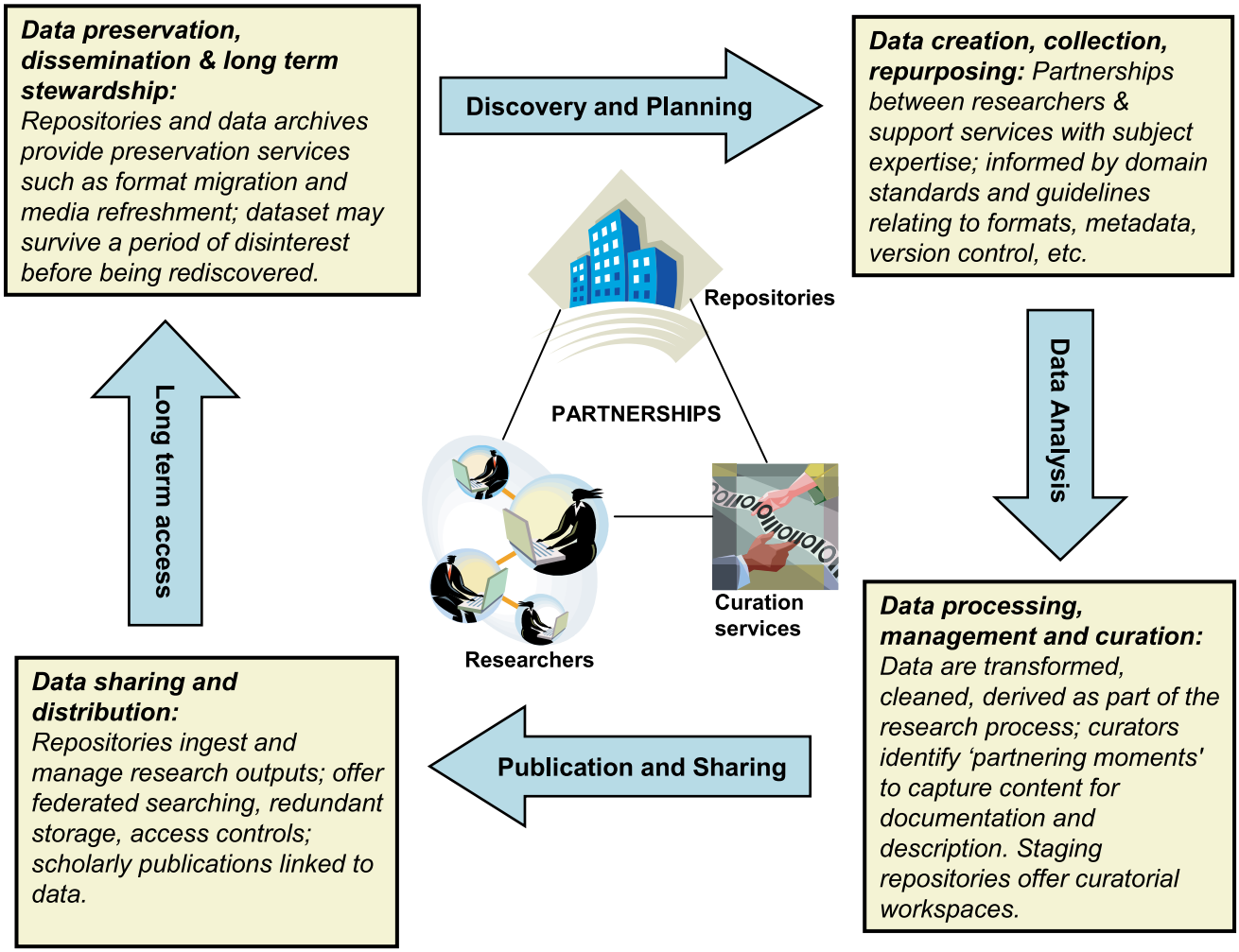


Figure 1: Data Sharing Continuum

does for the UK Data Archive, or they may require grant applicants to include a data sharing plan, as the Medical Research Council has been doing since 2007. Similarly, the Data Quality Seal of Approval, developed by DANS--Data Archiving and Networked Services—in the Netherlands, stakes out roles and responsibilities of different players for assuring data quality for data in repositories (Sesink, et al 2008). Intriguingly, they include users as part of the equation.

Our experience so far shows that even where academics are not interested in sharing their data publicly, they do recognise the importance of data management and are interested in the possibility of getting institutional support for improving current practice. The project has benefited from the input of its consultant, Digital Life Cycle Research & Consulting, with regard to the need for a life-cycle approach to data curation and the partnerships that could be forged throughout that life cycle (Green and Gutmann, 2007). One of the findings from such a perspective is that librarians – in their roles as either data librarians or repository managers – could find ways to move "upstream" in the research process, i.e. get involved in the pre-publishing stages where data is created and processed, rather than the usual librarian's comfort zone of dealing with post-published materials "downstream" (Gold, 2007). (See figure 2 below)

Identifying and describing the data management requirements of digital collections is a central part of understanding what roles and services are required for research data management. One of the additional deliverables taken on by the three DataShare partners for the second part of the project is to conduct data audits in partnership with academic departments using the tools and methodology developed by the Digital Curation Centre as part of the Data Audit Framework[6] development project. JISC funded the project in response to one of the many



Legend: Research lifecycle: blue arrows, Data lifecycle: yellow boxes

Figure 2: *Partnerships in the Data & Research Lifecycle (courtesy Digital Lifecycle Research & Consulting)*

recommendations in the *Dealing with Data* report:

> *A framework must be conceived to enable all Universities and colleges to carry out an audit of departmental data collections, awareness, policies and practice for data curation and preservation. (Lyon, 2007).*

Open data and open licenses

Open access repositories allow any user to access their content via the WWW as well as allowing other servers to access and harvest their metadata, e.g. Google or scholarly search engines. The following definition is from the Budapest Open Access Initiative (2002):

> *By 'open access' to this [scientific and scholarly journal] literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself.*

The more recent Open Knowledge Foundation definition equates open access with open content, rather than just research literature: "A piece of knowledge is open if you are free to use, reuse, and redistribute it"[7] (similar to the open source movement for software). This is not to be confused with the 'open data definition' used for copying personal data from one social networking site to another.[8]

As Peter Murray-Rust has stated, "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." (Poynder, 2008) In other words, for data to be useful, they need not only be read, but manipulated, re-used, re-coded, mined, and merged (or mashed, if you will) with other data. For a chemist like Murray-Rust who analyses chemical structures by mining chemical literature for tables, charts, images containing data about molecular structures, not only is a PDF document not sufficient to re-use the data embedded within, but the restrictions on the re-use of the content by publishers and even supposedly open access repositories are too stringent.

For this reason, and to generally encourage the proliferations of mashups on the Web, the open data community developed an "open data license" for data publishers (i.e. those responsible for making their data available) to set their data free. First, Science Commons, a project under the banner of Creative Commons that deals with licensing copyrighted materials, undertook the development of a protocol upon which any open data license would be based:

Science Commons' Protocol for Implementing Open Data[9]

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.

Following this, a number of players including law scholars at the University of Edinburgh were responsible for bringing about the Public Domain Dedication and License, which attempts to include wording that either waives IPR altogether or in cases where that is not legally possible, dedicates it to the public domain. Key concepts covered by the PDDL are:

- 'Converge on the public domain' by waiving all rights based on intellectual property
- Take into account "sui generis" database right (in European jurisdictions, e.g. Database Directive rights)
- Avoid attribution stacking (as the "attribution" norm is a burden when merging from highly numerous sources of data)

The Edinburgh DataShare repository offers an option to attach a PDDL to deposited datasets. Where depositors do not wish to freely give away their data, or are prevented from doing so based on agreements with subjects, funders, or research ethics boards, they may fill out a Rights statement field spelling out the terms of use, or fill out a metadata-only record where potential users are free to contact the depositor to request access.

Conclusion

Data management, curation and publishing are getting much attention globally and in the UK at present. There are a number of nagging problems that have not been solved over the years, such as the lack of career reward for publishing data as opposed to papers and the lack of career paths for 'data scientists' (National Science Board, 2005; Swan and Sheridan, 2008b). Related to this is the poor practice of data citation, especially where data are shared only informally between peers or downloaded from a website rather than obtained from an archive or repository with a complete metadata record. The scattered infrastructure for data curation across disciplines and institutions is being addressed in Australia through the ANDS national data service[10] and a feasibility study for a more modest shared research data service in the UK[11] . Canada has stepped forward with the Research Data Strategy Working Group to address the challenges surrounding the access and preservation of research data[12] . Two American Universities, Cornell and MIT, are pursuing ground-breaking, library-led data curation services for their users. These are described in sister articles in this issue of IQ. Perhaps the current critical mass of attention and

effort will help break through the remaining barriers that prevent data from being cared for, shared, used to develop new knowledge, and preserved as an essential part of the scholarly record.

It has been an exciting time to be involved in a project such as DataShare. We are tracking as many of these developments as we can keep up with on our website. We welcome any and all feedback.

* Contact: Robin Rice, EDINA and Data Library, University of Edinburgh, UK. E-mail: R.Rice@ed.ac.uk

References

(2001). Budapest Open Access Initiative, February 14, 2002, Budapest: Open Society Institute. http://www.soros.org/openaccess/read.shtml

Carrier, S., J. Dube and J. Greenberg (2007). The DRIADE project: Phased application profile development in support of open science. In Proc. Int'l Conf. on Dublin Core and Metadata Applications 2007. DCMI. http://www.dcmipubs.org/ojs/index.php/pubs/article/view/39/19

The Center for Research Libraries and OCLC (2007). Trustworthy Repositories Audit & Certification: Criteria and Checklist. Version 1.0. February 2007. OCLC.http://www.crl.edu/PDF/trac.pdf

The Digital Archiving Consultancy, The Bioinformatics Research Centre (University of Glasgow) and The National e-Science Centre (NeSC) (2005). Large-scale data sharing in the life sciences: Data standards, incentives, barriers and funding models (the Joint data standards study). http://www.mrc.ac.uk/Utilities/Documentrecord/index.htm?d=MRC002552

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

Gold, A. (2007). Cyberinfrastructure, Data, and Libraries, Part 2. Libraries and the Data Challenge: Roles and Actions for Libraries, D-Lib Magazine 13(9/10). http://www.dlib.org/dlib/september07/gold/09gold-pt2.html

Green A. and Gutmann, M. P. (2007) Building partnerships among social science researchers, institution-based repositories and domain specific data archives. OCLC Systems and Services, 23 (1), 35-53. http://deepblue.lib.umich.edu/handle/2027.42/41214 [Open Access version]

Heery, R. and Anderson, S. (2005). Digital repositories review. http://www.jisc.ac.uk/uploaded_documents/digital-repositories-review-2005.pdf

Heery, R. and Powell, A. (2006). Digital repositories roadmap: looking forward. Bath: UKOLN/Eduserv. http://www.ukoln.ac.uk/repositories/publications/roadmap-200604/

Humphrey, C.K., Estabrooks, C.A., Norris, J.R., Smith, J.E. and K.L. Hesketh (2000). Archivist on board: Contributions to the research team. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research 1(3). http://qualitative-research.net/fqs/fqs-eng.htm

Lyon L. (2007) Dealing with data: roles, responsibilities and relationships, Consultancy Report. June, 2007, Bath: UKOLN. http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report-final.pdf

Macdonald, S. (2008a). Data Visualisation Tools: Part 1 - Numeric Data in a Web 2.0 Environment. DISC-UK, January 2008. http://www.disc-uk.org/docs/Numeric_data_mashup.pdf

Macdonald, S. (2008b). Data Visualisation Tools: Part 2 - Spatial Data in a Web 2.0 Environment and Beyond. DISC-UK, September 2008. http://www.disc-uk.org/docs/spatial_data_mashup_V2.pdf

Martinez, L. (2008). The Data Documentation Initiative (DDI) and Institutional Repositories. DISC-UK, February 2008. http://www.disc-uk.org/docs/DDI_and_IRs.pdf

National Science Board (2005) Long-lived digital data collections: enabling research and education in the 21st Century. Washington, DC: National Science Foundation. http://www.nsf.gov/pubs/2005/nsb0540/

Poynder, R (2008). The Open Access Interviews: Peter Murray-Rust. Open and Shut [weblog]. January 21, 2008, http://poynder.blogspot.com/2008/01/open-access-interviews-peter-murray.html

Rice, R., Macdonald, S. and G. Hamilton (2008). Applying DC to Institutional Data Repositories. International Conference on Dublin Core and Metadata Applications (DC-2008,) 23-25 September, 2008, Berlin. http://dc2008.de/wp-content/uploads/2008/10/12_rice_poster.pdf

(2002). Reference Model for an Open Archival Information System (OAIS). Consultative Committee for Space Data Systems. January 2002.

Research Information Network. (2008) Stewardship of digital research data: a framework of principles and guidelines. January 2008, London: RIN.

Sesink, L., van Horik, R. and H. Harmsen (2008) Data

Seal of Approval: Quality guidelines for digital research data in the Netherlands. May, 2008, The Hague: Data Archiving and Networked Services (DANS). http://www.datasealofapproval.org

Seymour, R (2007). User based evidence for the requirements and functionality of a repository capable of managing licensed geospatial assets (public version), April, 2007, Edinburgh: EDINA. http://edina.ac.uk/projects/grade/FormalGISRepositoryFeedbackFinal.pdf

Swan, A. and S. Brown (2008a). To Share or not to Share: Publication and Quality Assurance of Research Data Outputs. June, 2008, London: Research Information Network. http://www.rin.ac.uk/data-publication

Swan, A. and S. Brown (2008b). The Skills, Role and Career Structure of Data Scientists: An Assessment of Current Practice and Future Needs. July, 2008, London: Joint Information Systems Committee. http://www.jisc.ac.uk/publications/publications/dataskillscareersfinalreport.aspx

Footnotes

1.http://www.disc-uk.org/datashare.html

2.http://ils.unc.edu/mrc/dryad

3. http://edina.ac.uk/projects/GAP_summary.html

4. This will be available on the project deliverables page - http://www.disc-uk.org/deliverables.html before March, 2009.

5. http://www.opendoar.org/tools/en/policies.php.

6. http://www.data-audit.eu/

7. http://www.opendefinition.org/

8. http://www.opendd.net/about.php

9. http://sciencecommons.org/projects/publishing/open-access-data-protocol/

10 http://ands.org.au/

11. http://www.ukrds.ac.uk/

12. http://cisti-icist.nrc-cnrc.gc.ca/media/press/rds_group_e.html