



## Project Document Cover Sheet

Project Information			
<b>Project Acronym</b>	DISC-UK DataShare (Data and Information Specialists Committee)		
<b>Project Title</b>	DISC-UK DataShare		
<b>Start Date</b>	March 2007	<b>End Date</b>	March 2009
<b>Lead Institution</b>	EDINA, University of Edinburgh		
<b>Project Directors</b>	Peter Burnhill (EDINA) and Mark Brown (University of Southampton)		
<b>Project Manager &amp; contact details</b>	Robin Rice, R.Rice@ed.ac.uk		
<b>Partner Institutions</b>	Universities of Oxford and Southampton		
<b>Project Web URL</b>	<a href="http://www.disc-uk.org/datashare.html">http://www.disc-uk.org/datashare.html</a>		
<b>Programme Name (and number)</b>	Repositories and Preservation programme (Startup and Enhancement strand)		
<b>Programme Manager</b>	Andrew McGregor		

Document Name			
<b>Document Title</b>	Final Report		
<b>Reporting Period</b>	for progress reports only		
<b>Author(s) &amp; project role</b>	Robin Rice, project manager		
<b>Date</b>	13 May, 2009	<b>Filename</b>	DataSharefinalreport.pdf
<b>URL</b>	In JISC IE Repository		
<b>Access</b>	<input type="checkbox"/> Project and JISC internal		<input checked="" type="checkbox"/> General dissemination

Document History		
Version	Date	Comments
1	15/4/2009	
2	13/5/2009	Corrections in acknowledgements, few other minor corrections

Project Acronym: DISC-UK DataShare  
Version: 1  
Contact: Robin Rice  
Date: 15 April, 2009

**JISC**

# **DISC-UK DataShare Project**

## **Final Report**

**Robin Rice, Project Manager**

**15 April, 2009**



# JISC Final Report

## Table of Contents

<b>TABLE OF CONTENTS</b> .....	<b>3</b>
<b>ACKNOWLEDGEMENTS</b> .....	<b>4</b>
<b>EXECUTIVE SUMMARY</b> .....	<b>5</b>
<b>BACKGROUND</b> .....	<b>6</b>
<b>AIMS AND OBJECTIVES</b> .....	<b>6</b>
<b>METHODOLOGY</b> .....	<b>7</b>
<b>IMPLEMENTATION</b> .....	<b>8</b>
<b>OUTPUTS AND RESULTS</b> .....	<b>9</b>
<b>OUTCOMES</b> .....	<b>9</b>
<b>CONCLUSIONS</b> .....	<b>10</b>
<b>IMPLICATIONS</b> .....	<b>10</b>
<b>RECOMMENDATIONS</b> .....	<b>11</b>
<b>REFERENCES</b> .....	<b>11</b>
<b>APPENDICES</b> .....	<b>12</b>
A. Q&A page from website	
B. Edinburgh partner report	
C. Oxford partner report	
D. Southampton partner report	
E. London School of Economics associate partner report	

## Acknowledgements

The project was funded from March 2007-March 2009 as part of JISC's Repositories and Preservation programme<sup>1</sup>, Repositories Enhancement strand. It was led by EDINA and Edinburgh University Data Library in partnership with the University of Oxford and the University of Southampton. The project built on the existing informal collaboration of UK data librarians and data managers who formed DISC-UK (Data Information Specialists Committee – UK)<sup>2</sup>.

### Project Team

Project Directors:

- Peter Burnhill - EDINA and Data Library, University of Edinburgh
- Mark Brown - University of Southampton

Project Manager:

Robin Rice - EDINA and Data Library, University of Edinburgh

Project Officers:

- Harry Gibbs - University of Southampton Library
- Stuart Macdonald - EDINA and Data Library, University of Edinburgh
- Cuna Ekmekcioglu – Information Services, University of Edinburgh
- Anne Donnelly - EDINA and Data Library, University of Edinburgh
- Jane Roberts - Oxford Data Library (Nuffield College)
- Luis Martinez Uribe – Oxford e-Research Centre

Repository Managers:

- Theo Andrew - Digital Library Division, University of Edinburgh
- Sally Rumsey - Oxford University Research Archive, Oxford University Library Services
- Neil Jefferies - IT Development & Strategy Team, Oxford University Library Services
- Wendy White - University of Southampton Libraries

Technical Staff:

- George Hamilton - EDINA National Data Centre
- Ben O'Steen - Oxford University Research Archive

We would like to acknowledge the contributions of our project consultant, Ann Green, Digital Lifecycle Research and Consulting, as well as our Associate Partner, the London School of Economics, particularly Tanvi Desai and Frances Shipsey. The Data Audit Framework Development team was helpful, in particular Sarah Jones, project manager, DCC/HATII, University of Glasgow. JISC Programme Manager, Andrew McGregor, was supportive and responsive throughout the project period.

---

<sup>1</sup> <http://www.jisc.ac.uk/whatwedo/programmes/reppres.aspx>

<sup>2</sup> <http://www.disc-uk.org/>

## Executive Summary

The DISC-UK DataShare Project was funded from March 2007-March 2009 as part of JISC's Repositories and Preservation programme, Repositories Enhancement strand. It was led by EDINA and Edinburgh University Data Library in partnership with the University of Oxford and the University of Southampton. The project built on the existing informal collaboration of UK data librarians and data managers who formed DISC-UK (Data Information Specialists Committee – UK).

This project has brought 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, and thus to explore new pathways to assist academics at our institutions who wish to share their data over the Internet.

The project's overall aim was 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.

With three institutions taking part plus the London School of Economics as an associate partner, a range of exemplars have emerged from the establishment of institutional data repositories and related services. Part of the variety in the exemplars is a result of the different repository platforms used by the three project partners: DSpace (Edinburgh DataShare), ePrints (e-Prints Soton) and Fedora (Oxford University Research Archive, ORA)--all open source software. LSE took another route and is using the distributed Dataverse repository network for data, linking to publications in LSE Research Online. Also, different approaches were taken in setting up the repositories. All three institutions had an existing, well-used institutional repository, but two chose to incorporate datasets within the same system as the publications, and one (Edinburgh DataShare) was a paired repository exclusively for datasets, designed to interoperate with the publications repository (Edinburgh Research Archive).

The approach took a major turn midway through the project when an apparent solution to the problem of lack of voluntary deposits arose, in the form of the advent of the Data Audit Framework. Edinburgh participated as a partner in the DAF Development project which created the methodology for the framework, and also won a bid to carry out its own DAF Implementation project. Later, the other two partners conducted their own versions of the data audit framework under the auspices of the DataShare project.

A number of scoping activities were carried about by the partners with the goal of informing repository enhancement as well as broader dissemination. These included a State-of-the-Art-Review to determine what had been learned by previous repository projects in the UK that had forayed into the data arena. This resulted in a list of benefits and barriers to deposit of datasets by researchers to inform our outreach activities. A Data Sharing Continuum diagram was developed to illustrate where the projects were aiming to fit into the curation landscape, and the range of curation steps that could be taken, from simple backup to online visualization. Later on, a specialized metadata schema was explored (Data Documentation Initiative or DDI) in terms of how it might be incorporated into repository systems, though repository development in this area was not taken up. Instead, a dataset application profile was developed based on qualified Dublin Core (dcterms). This was implemented in the Edinburgh DataShare repository and adapted by Southampton for their next release. The project wished to explore wider issues with open data and web publishing, and therefore produced two briefing papers to do with data mashups – on numeric data and geospatial data. Finally, the project staff and consultant distilled what it had learned in terms of policy development for data repositories in a training guide. A number of peer reviewed posters, papers, and articles were written by DISC-UK members about various aspects of the project during the period.

Key conclusions were that 1) Data management motivation is a better bottom-up driver for researchers than data sharing but is not sufficient to create culture change, 2) Data librarians, data managers and data scientists can help bridge communication between repository managers & researchers, and 3) IRs can improve impact of sharing data over the internet.

## Background

This project has brought 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, and thus to explore new pathways to assist academics at our institutions who wish to share their data over the Internet. This collaboration arises from an existing UK consortium of data support professionals working in departments and academic libraries in universities (Data Information Specialists Committee-UK), and builds on an international network with a tradition of data sharing and data archiving dating back to the 1960s in the social sciences.

By working together across four universities and internally with colleagues already engaged in managing open access repositories for e-prints, this partnership was in a unique position to introduce and test a new model of data sharing and archiving to UK research institutions. By supporting academics within the four partner institutions who wish to share datasets on which written research outputs are based, this network of institution-based data repositories develops a niche model for deposit of 'orphaned datasets' currently filled neither by centralised subject-domain data archives/centres/grids nor by e-print based institutional repositories (IRs).

The project – along with others funded simultaneously – has helped 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).

A JISC-commissioned report that was published during the first months of the project observed 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" (Lyon, 2007 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. The following year the Research Information Network examined the responsibilities of research institutions, funders, data managers, learned societies and publishers in turn (RIN, 2008).

'Open Data' has become 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 top-down drivers to open up publicly-funded research data to public availability (for example, OECD, 2007). However, evidence from surveys of researchers had shown that there were many reasons they did not want to share the data on which their research outputs were based (Pryor, 2007). The project aimed to explore the technical and environmental context of the open data movement while dealing pragmatically with the concerns of researchers about sharing their data.

## Aims and Objectives

The **project's overall aim** was 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.

### Objectives:

1. Build capacity of institutional repositories in UKHE to respond to the unique requirements of research datasets as a new 'document type'
2. Use a range of open source software repository solutions - Eprints, DSpace, Fedora - to provide exemplars and add-on tools for managing datasets as institutional repository items

3. Produce and disseminate findings - in cooperation with the Repositories Support Project (RSP) and the Repositories Research Team (RRT) - to inform library and repository managers about the organisational and technical issues associated with the deposit of research data
4. Work with the RSP, Digital Curation Centre (DCC) and others to identify training needs and solutions for increasing skills of information professionals in UKHE for managing research data

## Methodology

The general approach involved: a) being as transparent as possible in our decision-making so as to be informative to others considering taking the step of adding research data to their repositories, b) implementing simple rather than highly technical solutions whenever possible so that they could be adopted by other IRs, and c) educating ourselves through professional development opportunities and reporting back on the project blog, and by collecting relevant information on our website such as articles, blog posts, and reports on topics relevant to the project. Disseminating what we were learning about research data, potential depositors and about repositories was a prominent activity throughout the life of the project.

With three institutions taking part plus the London School of Economics as an associate partner, a range of exemplars have emerged from the establishment of institutional data repositories and related services. Part of the variety in the exemplars is a result of the different repository platforms used by the three project partners: DSpace (Edinburgh DataShare<sup>3</sup>), ePrints (e-Prints Soton<sup>4</sup>) and Fedora (Oxford University Research Archive, ORA<sup>5</sup>)--all open source software. LSE took another route and is using the distributed Dataverse<sup>6</sup> repository network for data, linking to publications in LSE Research Online<sup>7</sup>. Also, different approaches were taken in setting up the repositories. All three institutions had an existing, well-used institutional repository, but two chose to incorporate datasets within the same system as the publications, and one (Edinburgh DataShare) was a paired repository exclusively for datasets, designed to interoperate with the publications repository (Edinburgh Research Archive<sup>8</sup>).

Each partner followed its own methodological approach for development of its IR, with the project manager arranging regular telecon and annual face to face meetings to report and share progress. Digital Life Cycle Research & Consulting was retained to advise the partners due to experience both with research data and with institutional repositories – this also helped to give the project an international view, since the consultant was based in the US and was involved with universities there working on similar developments (Yale and Cornell).

The approach took a major turn midway through the project when an apparent solution to the problem of lack of voluntary deposits arose, in the form of the advent of the Data Audit Framework. Edinburgh participated as a partner in the DAF Development project<sup>9</sup> which created the methodology for the framework, and also won a bid to carry out its own DAF Implementation project. Later, the other two partners conducted their own versions of the data audit framework under the auspices of the DataShare project – making use of project funds freed up from the withdrawal of the fourth partner. This activity was fruitful in terms of getting closer to researchers (moving ‘upstream’ in the research process – Gold, 2007; Green and Gutman, 2007) and better understanding barriers to best practice in data management for potential depositors – raising a host of issues to be addressed within the institution before sharing through the repository could occur. Each of the partners continues to be involved with work that has resulted from the outcome of the DAF activity (Ekmekcioglu and Rice, 2008, Gibbs, 2009, Martinez-Urbe, 2009a).

---

<sup>3</sup> <http://datashare.edina.ac.uk/dspace/>

<sup>4</sup> <http://eprints.soton.ac.uk/>

<sup>5</sup> <http://ora.ouls.ox.ac.uk/>

<sup>6</sup> <http://dvn.iq.harvard.edu/dvn/>

<sup>7</sup> <http://eprints.lse.ac.uk/>

<sup>8</sup> <http://www.era.lib.ed.ac.uk/>

<sup>9</sup> <http://www.data-audit.eu/>

## Implementation

This covers the overall project; see the partner appendices (here and in each progress report) to see how IR development was implemented at each partner institution. (At Edinburgh a user requirements document was maintained by the developer on a project wiki, which each project member could add to.) The trio of a data librarian/manager, a repository manager and a technical expert/team working together at each partner site was the key to implementation. The long period without two of the three roles being in post at LSE was the reason for its withdrawal as a full partner.

Milestones in implementation for the project team as a whole were:

- April, 2007: Kick-off week, organised by project manager and hosted at each partner site with accompanying presentations from recently completed projects based at the partner institutions and brainstorming activities led by project consultant, Ann Green.
- August, 2007: *State of the Art Review* published as our first briefing paper, “to provide background information to inform the work of DataShare, to summarise and consolidate recent research and current policy relating to data sharing, and to identify knowledge gaps that may need to be addressed during the course of the project” (Gibbs, 2007).
- December, 2007: Four diagrams developed collaboratively for the project poster for the DCC conference in Washington, D.C. These lay the groundwork for future project dissemination through posters, flyers and conference papers. Each gives the project context in comparison with other contemporary work: diagram of partners’ experience, benefits and barriers of depositing data, data sharing continuum, and partnering in the data lifecycle (Rice, Green and Rumsey, 2007).
- April, May, July 2008: a critical mass of project team members were able to meet at the Open Repositories conference in Southampton, at the IASSIST conference in Stanford, California, and at the Repository Fringe event in Edinburgh. Project papers were given at each.
- February, 2008: A face-to-face meeting at EDINA in February brought project partners, the consultant, the programme manager, and local experts, including Charlotte Waelde from the AHRC Centre for Intellectual Property Law, together for two days of intensive discussion and talks. This helped with decision-making for each site on enhancements to their IRs.
- December, 2008: The occasion of the Edinburgh University Data Library’s 25<sup>th</sup> anniversary event (during the week of the DCC conference) brought the project team together in Edinburgh. A day-long project meeting at EDINA was followed by a public symposium of international speakers sharing their views on the historical and current nature of academic data support.<sup>10</sup>
- 2008: A string of invited papers, articles, talks and interviews were given by DISC-UK members—particularly Stuart Macdonald and Luis Martinez-Urbe—about the role and contributions of data librarians/professionals in the production of science.<sup>11</sup> The JISC-commissioned report on the career structures of data scientists helped to draw attention to the profession of data librarian, which was covered in the report (Swan and Sheridan, 2008).
- January, 2009: The project manager went on a study tour of Australian universities known for recent initiatives in institutional data support. As well as giving seminars about the DataShare project, she communicated her findings to the project team and others via the project blog.
- February, 2009: A meeting to discuss the findings of the UK Research Data Service feasibility study was held in London. Two of the three partners have signed up as pathfinder institutions if the UKRDS succeeds in receiving funding for the first phase. The work undertaken during the DataShare project helps to lay the groundwork for future activity under UKRDS in those two institutions (Oxford and Edinburgh).
- March, 2009: By the time of the project close, each partner had an operational instance of a data repository in operation, with complete policy, procedures and support in place (see partner reports in appendix). Additionally, they had a local body of knowledge and contacts built up from the data audit activity.

<sup>10</sup> See <http://datalib.ed.ac.uk/25anniversary/> for accounts of the event, photos and presentations.

<sup>11</sup> See <http://www.disc-uk.org/publications.html#pubs> for citations and links to these.

## Outputs and Results

A number of briefing papers were written by the project staff with the goal of informing repository enhancement as well as broader dissemination. A *State of the Art Review* was written early in the project to help us build on previous related work in the UK, particularly data-related prior JISC repository projects (Gibbs, 2007). Next, a specialized metadata schema for microdata (e.g. surveys) and aggregate data was explored (Data Documentation Initiative or DDI) in terms of how it might be incorporated into repository systems, though repository development in this area was not taken up since it was a workpackage for LSE (Martinez-Uribe, 2008). Instead, a dataset metadata profile was developed based on qualified Dublin Core (dcterms) and presented in a peer-reviewed poster at the Dublin Core conference (Rice, Macdonald and Hamilton, 2008). This was implemented in the Edinburgh DataShare repository<sup>12</sup> and adapted by Southampton for their demonstrator data repository (to be integrated with e-Prints Soton in the next software upgrade)<sup>13</sup>. Both are referenced in the forthcoming report from UKOLN on application profiles for scientific metadata (Ball, 2009).

To explore the Open Data movement and alternative methods of publishing data openly on the Internet, a pair of briefing papers were written exploring commercial and academic-based “mashups” of numeric and geospatial data that can be shared, combined, and visualised online, making use of new Internet protocols and markup languages such as the keyhole markup language (KML) used for Google maps (Macdonald, 2008a and 2008b).

We used various Web 2.0 tools on the project website which were curated over the course of the project. In particular we described deliverables in the blog<sup>14</sup>, which had a newsfeed, and used a shared bookmarking facility to annotate links to interesting discoveries in articles or blogs. A live tag cloud on the Collective Intelligence page<sup>15</sup> piped in keywords from the bookmarking site to create a dynamic bibliography of topics related to the project, which are also piped into the blog. One hundred and forty-two items were annotated and bookmarked. The site also maintains a complete collection of publications, posters and presentations from DISC-UK members and of seminal papers by others to do with data sharing and data support<sup>16</sup>. The website also has a Q&A page to explain basic concepts to do with the project to audiences with different tacit knowledge, e.g. librarians and researchers (attached as Appendix B).

Finally, the project staff and consultant distilled what they had learned in terms of policy development for data repositories in a training guide (Green, Macdonald, Rice, 2009). DISC-UK members will use it in a half-day workshop with data professionals at the forthcoming IASSIST conference in Tampere, Finland in May. A fuller discussion of the project as a whole and the issues encountered has been written for the *IASSIST Quarterly* (Rice, 2009). In the same addition is an account of the Oxford scoping study findings (Martinez-Uribe, 2009b).

*Please see Appendix A for specific deliverables completed during the final 6-month progress period. See also partner reports in the appendices for work completed during the final 6-month reporting period and in previous progress reports for full descriptions of IR enhancements and local work completed over the project period.*

## Outcomes

We feel that we have achieved the aims and objectives we set for ourselves, with the caveat that we have more local outreach to do to gain a critical mass of deposits in our respective repositories. Had the project been given three years as in the previous digital repositories programme, we might have had sufficient time to seed the repositories with more than just a few datasets.

---

<sup>12</sup> [http://www.disc-uk.org/docs/Edinburgh\\_DataShare\\_DC-schema1.pdf](http://www.disc-uk.org/docs/Edinburgh_DataShare_DC-schema1.pdf)

<sup>13</sup> [http://www.disc-uk.org/docs/sPrints\\_Soton\\_Metadata.pdf](http://www.disc-uk.org/docs/sPrints_Soton_Metadata.pdf)

<sup>14</sup> <http://jisc-datashare.blogspot.com/>

<sup>15</sup> <http://www.disc-uk.org/collective.html>

<sup>16</sup> <http://www.disc-uk.org/publications.html>

On our website we wrote the following intended outcomes, and we feel we have achieved these.

- Exemplars of the process, pitfalls and successful outcomes of setting up an institutional data repository service at each of the three institutions
- 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
- Enhancements to partners' IRs including testing for trusted repository status
- Technical watch on e-Research, VREs, Web 2.0 and related developments. Papers, presentations and online dissemination of collected knowledge

Our external evaluator will provide a complete summative evaluation of our work at the end of April.

## Conclusions

The following "lessons learned" were presented at the Digital Curation Practice, Promise and Prospects conference<sup>17</sup> by the project manager on 1-3 April, 2009 in Chapel Hill, North Carolina.

- 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<sup>18</sup>)
- Data librarians, data managers and data scientists can help bridge communication between repository managers & researchers
- 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!
- *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!*

## Implications

Implications for the community are considered in the section above.

In terms of development that could build on this work, the two briefing papers on data visualisation tools point the way to research use of such cloud or commercial online tools, including potential links with IRs. The future UKRDS (UK Research Data Service) could potentially build on our work by considering the role of institutional repositories alongside domain-specific repository provision for sharing data assets between research institutions. Since our development work involved only scattered "volunteer" depositors from various parts of our universities, questions about scaling up – both technical and in terms of service support (e.g. assisted deposit) remain, if repositories become a more widespread data sharing solution. This is not likely unless institutions choose to take more responsibility for stewardship of data assets produced by their members. For this reason, policy development for research data at research institutions is needed. Most preservation initiatives at Universities focus on library special collections (especially involving digitisation), published research

<sup>17</sup> <http://ils.unc.edu/digccurr2009/>

<sup>18</sup> [http://www.disc-uk.org/docs/data\\_sharing\\_continuum.pdf](http://www.disc-uk.org/docs/data_sharing_continuum.pdf)

outputs, and administrative records rather than data assets. In this sense, the Data Audit Framework is before its time, since researchers do not believe that institutions have a role to play in data stewardship and rightly feel they are on their own. We hope that JISC will take forward the scientific metadata application profile work from the feasibility study mentioned above and that the Edinburgh and Southampton work on Dublin Core metadata for datasets will make a contribution to that. The Dublin Core Metadata Initiative community has also started an international working group on metadata standards for scientific data.<sup>19</sup> We feel that DISC-UK has contributed to the new discourse about career paths for data professionals and will continue to participate via the DCC/RIN Research Data Management Forum, which we hope will recommend actions for training, education, professional development and certification for data librarians, data managers and data scientists.

## Recommendations

Please see Conclusions and Implications sections, above.

## References

- Ball, Alex (2009). *Scientific Data Application Profile Scoping Study report*. Bath: UKOLN (forthcoming).
- Ekmekcioglu, Ç. and Rice, R. (2009). *Edinburgh Data Audit Implementation Project: Final report*. Edinburgh: University of Edinburgh, January 2009. <http://ie-repository.jisc.ac.uk/304/>
- 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>
- Gibbs, Harry (2009). *Southampton data survey: our experience and lessons learned*. Southampton: University of Southampton, March, 2009. <http://ie-repository.jisc.ac.uk/304/>
- 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 Gutman, 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]
- Green, Macdonald, Rice (2009). *Policy-making for research data in repositories: A guide*. DISC-UK: May, 2009. <http://www.disc-uk.org/docs/guide.pdf>
- Heery, R. and Anderson, S. (2005). Digital repositories review. UKOLN, AHDS: 19 February, 2005. [http://www.jisc.ac.uk/uploaded\\_documents/digital-repositories-review-2005.pdf](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, April 2006. <http://www.ukoln.ac.uk/repositories/publications/roadmap-200604/>
- Lyon L. (2007). Dealing with data: roles, responsibilities and relationships, Consultancy Report. June, 2007, Bath: UKOLN. <http://www.jisc.ac.uk/publications/publications/dealingwithdatareportfinal.aspx>
- 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](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](http://www.disc-uk.org/docs/spatial_data_mashup_V2.pdf)

---

<sup>19</sup> <http://dublincore.org/groups/sam/>

- 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](http://www.disc-uk.org/docs/DDI_and_IRs.pdf)
- Martinez-Uribe, Luis (2009a). *Using the Data Audit Framework: an Oxford case study*. Oxford: University of Oxford, March 2009. <http://ie-repository.jisc.ac.uk/300/>
- Martinez-Uribe, L. (2009b). Digital repository services for managing research data: What do Oxford researchers need? *IASSIST Quarterly* 31:3-4, Fall, 2007, p. 29-34. [http://www.iassistdata.org/publications/iq/iq31/iqvol313\\_4martinez-uribe.pdf](http://www.iassistdata.org/publications/iq/iq31/iqvol313_4martinez-uribe.pdf)
- OECD (2007). *OECD Principles and Guidelines for Access to Research Data from Public Funding*. OECD. <http://www.oecd.org/dataoecd/9/61/38500813.pdf>
- Pryor, G. (2007). Project StORe: Making the connections for research. *OCLC Systems and Services*, 23 (1), 70-78. <http://www.emeraldinsight.com/Insight/viewContentItem.do?contentType=Article&contentId=1593463>. Open Access version:
- Rice, R., Green, A., and Rumsey, S. (2007). DISC-UK DataShare: Building capacity for institutional data repositories. *3rd International Digital Curation Centre conference*, Washington, DC, December 11-13 2007 [poster]. [http://www.dcc.ac.uk/events/dcc-2007/posters/DISC-UK\\_DataShare.pdf](http://www.dcc.ac.uk/events/dcc-2007/posters/DISC-UK_DataShare.pdf)
- Rice, R., Macdonald S. and G. Hamilton (2008). Applying DC to institutional data repositories. *International Conference on Dublin Core and Metadata Applications (DC-2008)* Berlin, 23-25 September, 2008 [poster]. [http://dc2008.de/wp-content/uploads/2008/10/12\\_rice\\_poster.pdf](http://dc2008.de/wp-content/uploads/2008/10/12_rice_poster.pdf)
- Rice, R. (2009). DISC-UK DataShare Project: Building exemplars for institutional data repositories in the UK. *IASSIST Quarterly* 31:3-4, Fall, 2007, p. 22-28. [http://www.iassistdata.org/publications/iq/iq31/iqvol313\\_4rice.pdf](http://www.iassistdata.org/publications/iq/iq31/iqvol313_4rice.pdf)
- RIN (2008). *Stewardship of digital research data: a framework of principles and guidelines*. January, 2008, London: Research Information Network. <http://www.rin.ac.uk/files/Research%20Data%20Principles%20and%20Guidelines%20full%20version%20-%20final.pdf>
- Swan, A. and S. Brown (2008). *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>

## Appendices

- A. Outputs and Deliverables: November 2008 – April 2009
- B. Q&A page from website
- C. Edinburgh partner report
- D. Oxford partner report
- E. Southampton partner report
- F. London School of Economics associate partner report

## Appendix A

### ***Deliverables and Milestones: November 2008 – April 2009<sup>20</sup>***

1. *November, 2008:* Stuart Macdonald presented [Open Data - Projects, Tools, Initiatives](#) at [Seminario Sobre Datasets Consorcio Madrono](#), Salon de Actos, Facultad de Ciencias Politicas y Sociologia - UNED, Madrid, 17 November 2008.
2. *November, 2008:* Robin Rice and Guy McGarva (EDINA) gave a two minute presentation at the [Innovation Fair](#) on *Geospatially Enabling DSpace Repositories: Standards-compliant ways to georeference items in DSpace* at the SPARC Conference on Digital Libraries in Baltimore, 18 November 2008.
3. *November, 2008:* Robin Rice gave a presentation on [Roles & Responsibilities for Data Curation: the Data Librarian](#) at the 2nd DCC/RIN workshop of the Research Data Management Forum in Manchester, 27 November 2008.
4. *December, 2008:* The DISC-UK project team met in person on 4th December at EDINA in Edinburgh (where the DCC conference was held earlier in the week).
5. *December, 2008:* A [Symposium on Institutional Data Services](#) was held as part of the 25th Anniversary Celebration of the Edinburgh University Data Library on 5 December. Peter Burnhill, Director of EDINA and the Data Library, presented a paper on the history of the Data Library. Other speakers were Charles (Chuck) Humphrey and Sheila Anderson plus an international panel.
6. *December, 2008:* Anne Donnelly joins the project team as DataShare Project Officer at Edinburgh.
7. *January, 2009:* The final report of the [Edinburgh Data Audit Framework Implementation Project](#) was submitted and deposited in the JISC Information Environment repository, including 5 case studies of audited units in the University. Cuna Ekmekcioglu, the project manager of that project is seconded to the DataShare project to assist the DataShare project manager and the local DAF steering committee to carry out its intended outcomes from the project, including development of training and guidance web pages for data management aimed at researchers and postgraduates. This work will continue into the summertime (after the end of the DataShare project).
8. *January, 2009:* The project manager made a number of study visits to Australian and New Zealand institutions engaged in developing services and policies for data management. Her impressions are being recorded in a series of posts on the [DataShare blog](#): the series is called [Data Walkabout](#). Copies of her presentations at seminars at the [University of Sydney](#) and [Monash University](#) (plus a [podcast](#) at the latter) have been posted.

---

<sup>20</sup> Since the final report uses a different template from previous progress reports, this appendix has been added in order to give a more complete account of work completed during the final 6-month project period.

9. *February, 2009:* Harry Gibbs wrote a [blog post](#) summarising the event, "The significance of data management for social survey research," hosted by the Economic and Social Data Service at the University of Essex.
10. *February, 2009:* Anne Donnelly wrote a [blog post](#) describing the *JISC Start Up & Enhancement Projects Training Event: Embedding Repositories*, University of Lincoln, 10th February 2009.
11. *February, 2009:* Stuart Macdonald gave an invited presentation, [The strange case of the local data librarian - a peculiarly Edinburgh perspective!](#) at the Economic and Social Data Service training event, [ESDS - What's in it for Librarians?](#) on 13 Feb, 2009.
12. *February, 2009:* George Hamilton wrote a [blog post](#) of his impressions of attending the *JISC Developer Happiness Days* in London, 9-13 February.
13. *March, 2009:* 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*. A [blog post](#) by Ben O'Steen discusses handling tabular data in *DataBank*, and [another post](#) on the *DataShare* blog by Luis Martinez Uribe gives an update of recent progress at Oxford.
14. *March, 2009:* Oxford and Southampton partners have completed their Data Audit Framework deliverables and published the following reports: [Using the Data Audit Framework: an Oxford Case Study](#), and [Southampton Data Survey: Our Experience and Lessons Learned](#).
15. *March, 2009:* A prototype ePrints 3.1 repository has been created at the University of Southampton. This has enabled the deposit of research data, and associated metadata, under the new item type 'Dataset'. It has also provided an opportunity to carry out usability testing on the deposit process and metadata schema. The repository is not intended to be made public and instead will act as an internal demonstrator to encourage and facilitate data deposit. When the institutional repository, [ePrints Soton](#), is upgraded to 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.
16. *March, 2009:* Metadata Schema. [DataShare Metadata Schema for ePrints Soton](#), by Harry Gibbs [PDF].
17. *March, 2009:* The Questions and Answers page has been completed on the project website by the project manager; the purpose is to explain basic concepts to do with the project to audiences with different tacit knowledge, e.g. librarians and researchers.
18. *April, 2009:* One hundred and forty-two "[faves](#)" (annotated social bookmarks) were written by project staff over the project period on topics relevant to the project. The [dynamic tag cloud](#) on the DISC-UK collective intelligence page displays links to the bookmarks by keyword tags and the annotations have also been piped into the project [blog](#).

19. *April, 2009*: The project manager gave an invited paper - [Lessons Learned from the DISC-UK DataShare and Data Audit Framework Implementation Projects](#) at the Digital Curation Practice, Promise and Prospects (DIGCCUR) [conference](#) in Chapel Hill, North Carolina, 1-3 April.
20. *April, 2009*: Final project report submitted to JISC including partner appendices.
21. *May, 2009*: The project staff and consultant distilled what they had learned in terms of policy development for data repositories in a [training guide](#) to be published this month. DISC-UK members will use it in a half-day [workshop](#) with data professionals at the forthcoming IASSIST conference in Tampere, Finland this month.
22. *May, 2009*: Two papers were contributed to the most recent edition of [IASSIST Quarterly](#): the project manager summarised the work of the DataShare Project; Luis Martinez-Uribe reported on the requirements gathering exercise on researchers' needs at Oxford.
23. *May, 2009*: An external evaluation of the project is being undertaken by Sheila Anderson, Kings College London, and will be submitted to JISC at the end of May.

## Appendix B

### Questions and Answers

This page<sup>21</sup> aims to explain basic concepts important to the project:

1. Q. What are digital repositories?
2. Q. What are institutional repositories?
3. Q. What does the project mean by 'data'?
4. Q. What is a dataset / data set?
5. Q. What is a data library?
6. Q. What is secondary analysis?
7. Q. What is open access?
8. Q. What is open data?
9. Q. What is metadata for?
10. Q. What is the value of data repositories?

- 
1. According to the *Digital Repositories Review*, (Heery and Anderson, 2005) a **repository** is differentiated from other digital collections by the following characteristics:
    - o content is deposited in a repository, whether by the content creator, owner or third party
    - o the repository architecture manages content as well as metadata
    - o the repository offers a minimum set of basic services e.g. put, get, search, access control
    - o the repository must be sustainable and trusted, well-supported and well-managed.
  2. 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 or domain-specific repositories, such as Arxiv (for Physics papers) and Jorum (for learning objects).
  3. By **data**, we do not mean a synonym for information. We mean research data, that which is collected, observed, or created, for purposes of analysing to produce original research results. This differs from what is commonly called research outputs, which are the peer reviewed, published papers/articles/books/presentations that are produced as a result of data analysis. Research data may be created in tabular, statistical, numeric, geospatial, image, multimedia or other formats.
  4. **Datasets** (or data sets) are a group of data files--usually numeric or encoded--along with the documentation files (such as a codebook, technical or methodology report, data dictionary) which explain their production or use. Generally a dataset is un-useable for sound analysis by a second party unless it is well documented.

---

<sup>21</sup> <http://www.disc-uk.org/ganda.html>

5. A **data library** refers to both the content and the services that foster use of collections of numeric and/or geospatial data sets for secondary use in research. (Wikipedia, 2007, [http://en.wikipedia.org/wiki/Data\\_libraries](http://en.wikipedia.org/wiki/Data_libraries)).
  
6. **Secondary analysis** is common in the social sciences, whenever a data source is used that was collected by someone other than the researcher involved. It's a method that saves time and expense, as long as the data used is reliable and fit for purpose. Large-scale datasets such as surveys are frequently not 'exhausted' by the original data collector and are therefore a rich source for secondary analysis.
  
7. **Open Access** means access to material via the Internet in such a way that the material is free for all users to read and use. (Wikipedia, 2009, [http://en.wikipedia.org/wiki/Open\\_access](http://en.wikipedia.org/wiki/Open_access)) The open access publishing movement was started by the [Budapest Open Access Initiative](#) and its signatories in February 2002.
  
8. **Open Data** is a philosophy and practice requiring that certain data are freely available to everyone, without restrictions from copyright, patents or other mechanisms of control. It has a similar ethos to a number of other "Open" movements and communities such as open source, and open access and open content. (Wikipedia, 2009, [http://en.wikipedia.org/wiki/Open\\_data](http://en.wikipedia.org/wiki/Open_data)) The OECD and other international bodies have endorsed the concept that publicly funded research data should be made freely available to the public. "Mashups" or web services are often based on combining various sources of open data which would not be possible to create if restrictions on their use were in place.
  
9. **Metadata** means information about a data item in the repository, including descriptive metadata such as title, and administrative metadata such as date of submission. Metadata in the repository can be searched by Google and specialist search engines and is important for tracking provenance of the dataset.
  
10. Research has shown that researchers consider access to a publication's primary source data a significant advantage to their own research. **Digital repositories facilitate the trend towards global research collaboration.** By sharing research data researchers enable both secondary analysis and the exploration of topics not envisioned by the initial investigator.

## Advocacy and Promotion

The Data Library celebrated its 25th anniversary year of supporting staff and students in the discovery, access, use and management of research datasets by holding a two-part day event. On 5 December 2008, following on from the Digital Curation Centre international conference held in Edinburgh that week, friends and colleagues from near and far gathered in the afternoon at the EDINA offices to celebrate the milestone with short speeches and cutting the cake, along with food, drink and toasts. Earlier in the day, about 45 professionals and academics gathered to consider the significant changes that the Data Library has been part of and might look forward to in the future:

*Symposium on Institutional Data Services Pollock Halls, 5th December, 10.00-15.00*

- How do data libraries adapt to users' evolving needs as technology, research topics and methods advance?
- How do institutional data services fit best in the national and international network of data services?
- What models besides data libraries are there for providing data-related services including data curation?
- What national initiatives around data services and data sharing are worthy of attention in the UK?

Three speakers -- Sheila Anderson (Kings College London), Chuck Humphrey (University of Alberta) and Peter Burnhill (EDINA and Data Library) -- shared their thoughts and engaged with the audience in discussion. Following a 'networking lunch', an international panel presented top themes from four countries: Ann Green, USA; Henk Harmsen, the Netherlands; Robin Rice, UK; and Andrew Treloar, Australia. A paper was written by Peter Burnhill to mark the occasion: [Edinburgh University Data Library: The First 25 Years](#).

[Slides and photos from the event](#) were posted on the Data Library website and it was written up in the IS newsletter for University staff--[BITS](#), as well as on the [DataShare blog](#) and as an EDINA news item. The event was supported by JISC through the DataShare project.

Anne Donnelly commenced as part-time DataShare Project Officer in mid-December and took on the task of drawing up an advocacy plan for actively promoting the repository to staff at the University. A promotional statement on the benefits of depositing data in *Edinburgh DataShare*, with definitions of terms and repository policies, now features on the [About Edinburgh DataShare](#) web page. The pilot project and Data Audit Framework activity identified several members of academic and support staff from across the Colleges who it is hoped may be able to give guidance and support in promoting DataShare to researchers in their respective schools. There has nevertheless been some spontaneous interest in the repository; earlier this year a dataset and supporting documentation was deposited by the School of Geosciences and a more recent expression of interest in doing so is being actively followed up. There are currently seven unique datasets in the

repository. Further promotional activity, aimed at raising awareness of the repository and encouraging the deposit of datasets by research staff across the University, is currently being planned.

In February 2009 the University's Senate passed an [Open Access Publications Policy](#) whereby, from January 2010, researchers will be formally required to deposit their research outputs in the University's closed Publication Repository and/or the open access Edinburgh Research Archive. Although it appears at the moment that 'research data issues' are being considered separately from research outputs, the formal embedding of this activity in academic practice at Edinburgh is a welcome development.

## Policy and Strategy

All of the activity under this heading in this period is crucial to the bedding down of the project into a full service sustained by the Data Library team as part of the University of Edinburgh Information Services.

The policies for *Edinburgh DataShare* were finalised in March and the public OpenDOAR record was modified, including removal of the term "pilot" from the description, indicating a full service. Using the OpenDOAR tool, policies for Metadata and Data re-use, Content types, Submissions, and Preservation were written and are recorded on the DataShare policy page (<http://datalib.ed.ac.uk/DataShare/policies.html>), in addition to a copy of the Depositor Agreement (license between depositor and the repository) written earlier which is incorporated in the DSpace software. Advanced preservation planning or procedures have not been developed yet, but the list of supported and "known" filetypes was modified in DSpace to help guide depositors to "future-proof" their deposited items.

Regular meetings with the Digital Library team within Information Services have been established, to inform one another of developments and to collaborate on areas of common concern (such as upgrading DSpace, presenting a compatible look and feel between the publications and data repositories, and developing usage statistics within DSpace). The new University requirements for deposit of publications mentioned in the Advocacy section above are raising questions about support for datasets, so it is important that advocacy undertaken for the publications repository by the Digital Library team is in a position to refer users to the *Edinburgh DataShare* service.

Much of the follow-up activity from the Edinburgh Implementation project of the Data Audit Framework (April-November 2008) is relevant to this project, and indeed was bid for and directed by the DISC-UK DataShare project manager. The project manager who carried out the audits and reporting for Edinburgh, Cuna Ekmekcioglu, from the Research Computing team, was seconded to the DataShare project for the remaining four months, since the desired outcomes identified by the DAF activity are compatible with the aims

of the DataShare project. This work will continue under the guidance of the local DAF steering committee, and includes the development of University web pages on guidance for data management, University policy development on data management, and service development including training courses for staff and postgraduates and a gap analysis. The DAF committee is broad-based and includes members from Research Computing, the Library, the University Archives, Records Management, Digital Curation Centre (DCC), Edinburgh Research Information (ERI), Brain Imaging Research Centre and the MRC Human Genetics Unit, as well as EDINA and Data Library.

The UK Research Data Service Feasibility Study has given its final report to HEFCE in the hopes of obtaining start-up funding for a "pathfinder" service as a transitional beginning to a fully-fledged shared data service. The Vice Principal of Knowledge Management, Chief Information Officer & Librarian, University of Edinburgh has put forward the University to be one of the pathfinder institutions, should funding commence. He has endorsed the continuing work of the (previous) DAF steering committee mentioned above in anticipation of this future activity and until a formal data working group is formed within the University. Cuna Ekmekcioglu will continue to work with Data Library staff on data management web pages and related activities through the summer, as part of her regular job in Research Computing.

## **Technical and Metadata Enhancements to IRs**

### **Upgrade to DSpace 1.5.1**

The DSpace repository was upgraded to the latest version 1.5.1.

### **Statistics**

Statistical functionality was lost with the upgrade from version 1.4.2 JSPUI to 1.5 Manakin client. A patch to DSpace version 1.5.1 allowed for statistics to again be gathered and displayed to admin users. An extension to DSpace to record bitstream downloads, which the standard installation does not support, implemented previously was ported to the new Manakin client.

### **Time Period**

A bespoke time period control was added to the DSpace ingest interface. The new control allows for a time/date period range to be recorded for a metadata item. The value is encoded and stored using the W3CDTF profile of ISO 8601.

### **Citation**

The citation value of an item is now automatically built using other metadata values when an item is created. The format of the value was agreed with other DataShare partners.

## **Spatial Coverage Auto-complete with GeoNames**

A simple geonames look up has been implemented in DSpace for populating the spatial coverage field.<sup>1</sup> The user is presented with an optional country drop-down box that lists all countries in the world and a second free text field. As the user types in the text field a drop down appears listing all possible completions for the text and country chosen (if a country is selected). Once the user selects an entry from the drop-down box, the place and country (if selected) are stored in the spatial coverage field. Currently, no co-ordinates are stored. This extension uses the GeoNames webservice for building the completion list.

## **Download All option**

A repository item containing data is likely to contain more than one bitstream. As a convenience to the user an option was added to the DSpace repository to be able to download all bitstreams in the item. The bitstreams, together with the license(s), are packaged in a zip file, on demand, and presented to the user for download.

## **Anti-Virus Checking**

All bitstreams uploaded with a new item to DSpace are now checked for viruses using the popular open source anti-virus software ClamAv. Any attempt to upload an infected bitstream will be rejected by the system. The virus database is automatically updated hourly.

## **New DSpace theme**

A new DSpace theme was created for the DataShare repository to give the repository a modern look-and-feel and to comply with the university's new branding. The opportunity was taken to improve for the experience for users browsing DataShare with particular attention to the view item page. A screenshot of the repository home page is appended. The location of the repository is <http://datashare.edina.ac.uk/dspace/>.

## **Assessing Impact**

As a start-up repository, two years is not a lot of time to assess the impact of *Edinburgh DataShare*. More outreach is needed into the research community to find more potential depositors; it was difficult to do this before the repository was a fully built service and before all the policies and features were in place. Current events including the Open Access Research Publications Policy and the UKRDS may have a role in increasing the importance for Edinburgh to have a place for research datasets to be deposited and shared.

---

<sup>1</sup> This was presented at the Innovation Fair of the SPARC Digital Repositories Meeting, Baltimore, Maryland, 17-18 Nov. 2008, along with an alternative implementation by the EDINA ShareGeo repository. See McGarva and Rice, <http://www.arl.org/sparc/meetings/ir08/innofair.shtml>.

The two projects - DataShare and DAF - have provided a tremendous focus for local collaboration for colleagues within and outwith Information Services (IS) who share concerns about curation of University data assets. In turn this has increased the visibility of the Data Library, which is a small service/team within IS--despite the size and reputation of EDINA as a national data centre.

Certainly the impact on the Data Library has been major--enabling very rich professional development of staff to learn and explore related topics of expertise, meet and network with others involved with similar work in the UK and abroad, and to garner the technical support to develop an entirely new online service to accompany the traditional data library service.

The ongoing follow-up work of developing data management web pages, recommendations for university policy, training and other services will hopefully create a large impact in the future. It is our hope to help the University meet its mission of "the creation, dissemination and curation of knowledge."

The screenshot shows the Edinburgh DataShare website. At the top left is the University of Edinburgh logo. The main header features the text 'THE UNIVERSITY OF EDINBURGH INFORMATION SERVICES' and 'Edinburgh DataShare'. On the right, there are links for 'University Homepage', 'IS Homepage', and 'Contacts'. Below the header, there is a navigation bar with 'Edinburgh DataShare >' and 'Contact Us | Login'.

The main content area is divided into three columns:

- Left Column:** Contains a search box for 'Search Edinburgh DataShare' with a 'Go' button, an 'Advanced Search' link, a 'Browse' section with a list of categories (Communities & Collections, By Issue Date, Data Creators, Titles, Subjects), and a 'My Account' section with 'Login' and 'Register' links.
- Middle Column:** Features a 'Welcome to Edinburgh DataShare' message, a paragraph describing the repository, a 'Search' section with a search box and 'Go' button, and a 'Communities in Edinburgh DataShare' section listing 'Information Services (IS)', 'School of GeoSciences', and 'School of History, Classics and Archaeology'.
- Right Column:** Includes a 'Links' section with links to 'About Edinburgh DataShare', 'Data Library Home', 'Edinburgh Research Archive', 'DISC-UK DataShare Project', and 'DataShare Blog'. It also has a 'Latest Items' section with an 'RSS Feed' icon and a list of recent items, including 'Carstairs deprivation scores by CATT2, 1981, 1991, 2001' and 'Refractive indices (500-3500 cm-) and emissivity (600-3350 cm-1) of pure water and seawater'.

At the bottom, there are logos for 'W3C XHTML 1.0' and 'W3C CSS', and a large 'JISC Repository Net' logo.

## Oxford DataShare Report - March 2009

### Advocacy and Promotion

The *Scoping Digital Repository Services for Research Data Management* project has been working in collaboration with the Oxford DataShare partners and has been involved in advocacy and promotion through a range of activities.

In the last months the project conducted a consultation with service units (Library, Computing Services, Research Services etc) across Oxford to find out about the *data management services* they offer. A framework<sup>1</sup> of research data management and curation services was created as part of this exercise. This consultation was complemented by a workshop, in October, which presented examples of the data services provided in Oxford and elsewhere and served to discuss the roles of service units in Oxford to support researchers.

In November Luis Martinez-Urbe participated in a data repositories seminar<sup>2</sup> organized by a consortium of academic libraries in Madrid. He presented a poster at the International Digital Curation Conference in December, and delivered a staff development seminar for librarians in Oxford.

The Data Audit Framework project in Oxford took place between August 2008 and March 2009. Two research groups participated in this work: the Young Lives project and the Cardiac Mechano-Electric Feedback Group. In addition to this, a set of data resources currently published on the University of Oxford website were identified and information about these data was compiled using the DAF register. A report has been produced that documents the data management practices of these groups mapping them to the DCC Curation Lifecycle Model, presents some of the data resources available on the Oxford website and discusses the experience of using the methodology.

### Policy and Strategy

The ICT Sub-committee of the Planning and Resource Allocation Committee (PRAC) of the University has produced a statement expressing the commitment to the curation of research data, in which they recognize the need to support researchers in accordance with the requirements of research funders. This statement has yet not been made publicly available.

### Technical and Metadata Enhancements to IRs

#### DataBank

A separate instance of Fedora has been set up which will act as a new repository running parallel to ORA. It will hold research data and has been named DataBank. We are anticipating that DataBank will be a store for 'long tail' data i.e. data not held in other locations within or external to Oxford University, and which does not comprise vast grid or similar datasets. In the first instance DataBank will not be directly

<sup>1</sup> <http://oxdrrc.blogspot.com/2008/12/research-data-management-services.html>

<sup>2</sup> <http://www.esds.ac.uk/international/news/madrid.asp>

accessible: access will be via digital objects held in ORA. The first content to be added to DataBank will be a collection of audio files produced as part of research in phonetics. Each file is a sample of a particular speech pattern. There are about 1,000 files in total, most lasting about 20 seconds, a small number lasting about 15 minutes. Access to these files will be via publications held in ORA which reference the files. It will not be possible at this stage to search for the files themselves. When testing has reached the point where we are satisfied with the performance, this instance of Fedora will be connected to the SUN Honeycomb storage layer where the files will ultimately be stored. Fedora will provide the digital object management layer and access will be via a web interface (initially ORA).

### Tabular data methodology

Documentation has been written giving details of a methodology for handling tabular data. It includes details of how to ingest tabular data into the repository, characterisation of the data, how to handle it once it has been ingested and how to access it. Storage of such data is with a view to preserving access. The documentation also includes a content model for tabular data for both Fedora and for BagIt<sup>3</sup> with RDF (Resource Description Framework). It is available at <http://oxfordrepo.blogspot.com/>.

### Assessing Impact

The interest from other universities worldwide on what Oxford is doing in the area of research data management and curation is clear. Luis Martinez-Uribe's blog, which reports the outputs of the projects and other relevant activities, has around 70 subscribers and visits from 64 countries.

Oxford's contribution, as one of four case study sites, to the UK Research Data Service (UKRDS) feasibility study, helped inform the planning of the next phase of the service's activities.

### Exit and Sustainability

The University of Oxford recognizes the need to address the management and long-term curation of research data. This recognition is expressed in the PRAC statement concerning management of research data at Oxford, and the identification of this activity as a key strategic priority in the University of Oxford ICT Strategic Plan<sup>4</sup>.

To continue the work in this area, a proposal has been submitted against the JISC Information Environment and e-Research Call for a project that seeks to implement elements of digital curation to address the data challenges of two experimental research groups in Oxford.

DataBank and the idea of data repositories will be included in the Digital Preservation Strategy currently being prepared for Oxford University Library Services, with the aim of firmly embedding such services within the institution.

---

<sup>3</sup> "a hierarchical file package format for the exchange of generalized digital content"  
<http://www.cdlib.org/inside/diglib/bagit/bagitspec.html>

<sup>4</sup> The University of Oxford ICT Strategic Plan:  
<http://www.ict.ox.ac.uk/strategy/plan/plan.xml.ID=S5#P262>

A number of other projects are also using or planning to use the DataBank repository, reusing elements of the DataShare work in terms of data modelling and description and helping to build a corpus of material.

In particular, the JISC-funded BID project aims to harvest metadata from the Oxford VLE (Sakai-based) and the e-Science Grid (SRB-based) to provide a framework for linking research data, research outputs and learning objects. This metadata will now be hosted within the DataBank repository.

A number of other project bids are in the pipeline where a data repository is a key enabling factor for the activities of the project.

Further to Oxford's contribution to the UKRDS feasibility study, the proposed next phase is the development of a Pathfinder service with a small group of key stakeholders, including the University of Oxford. The Pathfinder phase requires a commitment to long-term service provision and it is partly within the context of the UKRDS that the outcomes from these data management and curation related projects will be sustained.

## **DataShare: Southampton Final Progress Report March 2009**

### **Advocacy & Promotion**

During this final reporting period, advocacy work at Southampton centred around the Data Survey. Since October, the Data Audit evolved from a partial audit of data holdings to a survey of data types and data management processes. At the same time, the survey was widened to cover all six Divisions of the School of Social Sciences. It was felt that this broader approach would produce more interesting results and be more likely to uncover datasets suitable for deposit in ePrints Soton.

The partnership with social science Research Assistant, Teresa McGowan, has proved positive in terms of building relationships between the School and the Library. It is also felt that Teresa's good standing and contacts within the School contributed to the survey's pleasing participation rate. The methods used and issues raised in the Data Survey are detailed in the document entitled, 'Southampton Data Survey: Our Experience and Lessons Learned'.

Following the survey, informal discussions were held with five researchers about the nature of their data, storage and dissemination of that data and the possibility of deposit with ePrints Soton. A fifth researcher has also expressed an interest in discussing the preservation of his data in the near future.

To date this has resulted in the deposit of one dataset from a researcher who is interested in depositing further datasets once permission from collaborators is gained. A second researcher has agreed to deposit two datasets after Easter. These will be embargoed until 5 years after the date of collection. A third researcher is keen to deposit a well documented dataset but unfortunately the files are missing at present. The latter is an interesting case that highlighted the importance of careful data management and reliable storage methods.

A fourth researcher who has recently made some of his data available on open access via the Dataverse Network Project<sup>1</sup>, is keen to explore the possibility of disseminating his datasets through ePrints Soton. Pressure of his teaching load has prevented this so far. The fifth researcher was interested in procedures for storing and preserving data in physical formats to meet funder requirements. This raises further interesting questions for the University.

Further contact has been made with the Professor of Film Studies who is developing a database of German-Speaking Emigres in British Cinema. We are currently in the process of gathering the necessary metadata to complete the deposit.

---

<sup>1</sup> The Dataverse Network Project is housed at the Institute for Quantitative Social Science at Harvard University: <http://thedata.org/>

Finally, the Repository Manager has been contacted this month by a researcher in the School of Geography who is required to provide public access to a database as part of a NERC funded Programme. We are now in the process of arranging the deposit and expect to have the data publicly available by the Autumn, as the funder requires.

### **Strategy and Policy**

The shift in focus away from issues of open access towards those of data management and data sharing proved fruitful. It is now understood that researchers are unlikely to respond well when asked, in isolation, about making their data public. However, when approached in the context of their research and the data lifecycle, researchers are more open to the concept of data sharing.

The Data Survey demonstrates this in two respects. Firstly, the survey achieved significant engagement with the issues around data management, sharing and preservation. The survey's online questionnaire received a response rate of 38% of School researchers and almost half of those who reported that they held research data volunteered to take part in a follow-up interview.

Secondly, the survey revealed four researchers with an interest in depositing data in ePrints Soton. All are open to sharing in principle, although in some cases it will be necessary to restrict access to their data either for a specified period or indefinitely, for reasons of participant confidentiality. For one researcher, lack of storage space was the main motivation for deposit but she was also keen to disseminate her work with the hope of making contact with other researchers in the field.

### **Technical & Metadata Enhancements to IRs**

Owing to circumstances beyond DataShare's control, the upgrade of ePrints Soton to ePrints 3.1 is no longer scheduled to take place before the project end date. In view of this, it was agreed to develop the 'DataShare Prototype', an internal demonstrator ePrints 3.1 repository that can accept only datasets and will be absorbed into ePrints Soton after the upgrade. The Project Officer worked with the ePrints Soton Developer and ePrints Services to implement the metadata requirements discussed in the October Progress Report. Figure 1 (overleaf) shows the 'Details' deposit page of the DataShare Prototype.

Using the idea developed by Edinburgh DataShare, ePrints Services developed the Geonames function for the 'Spatial Coverage' field which allows depositors to autocomplete place names and country names from a drop-down list. When a geographical location is selected, ePrints stores (but does not display) longitude and latitude data from Geonames which is used to display locations in Google Maps. An example of this is shown in Figure 2. ePrints Services intends to develop this functionality and make the tool available to the wider ePrints community.

The Project Officer sat with a researcher during the deposit process and used this as an opportunity to carry out usability testing and to evaluate the metadata schema. This resulted in a number of small modifications being made to the Prototype, as well as comments being fed back to the developers of ePrints 3.1.

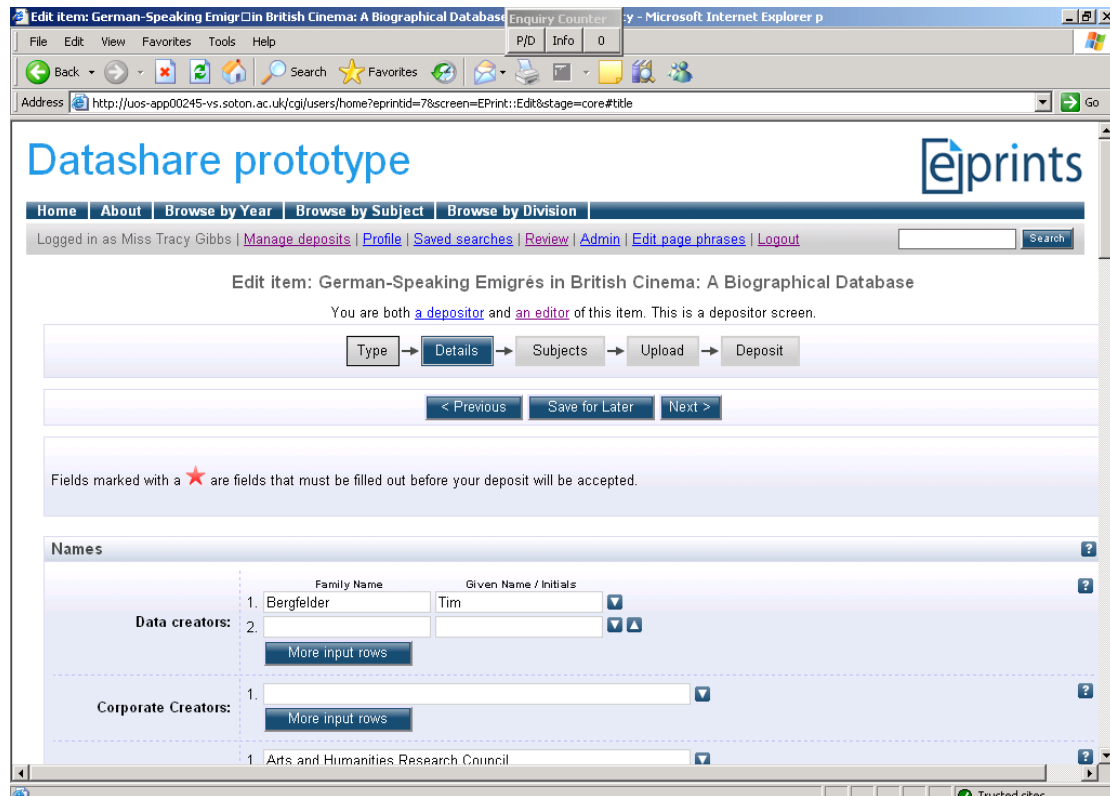


Figure 1: 'Details' deposit page of the DataShare Prototype

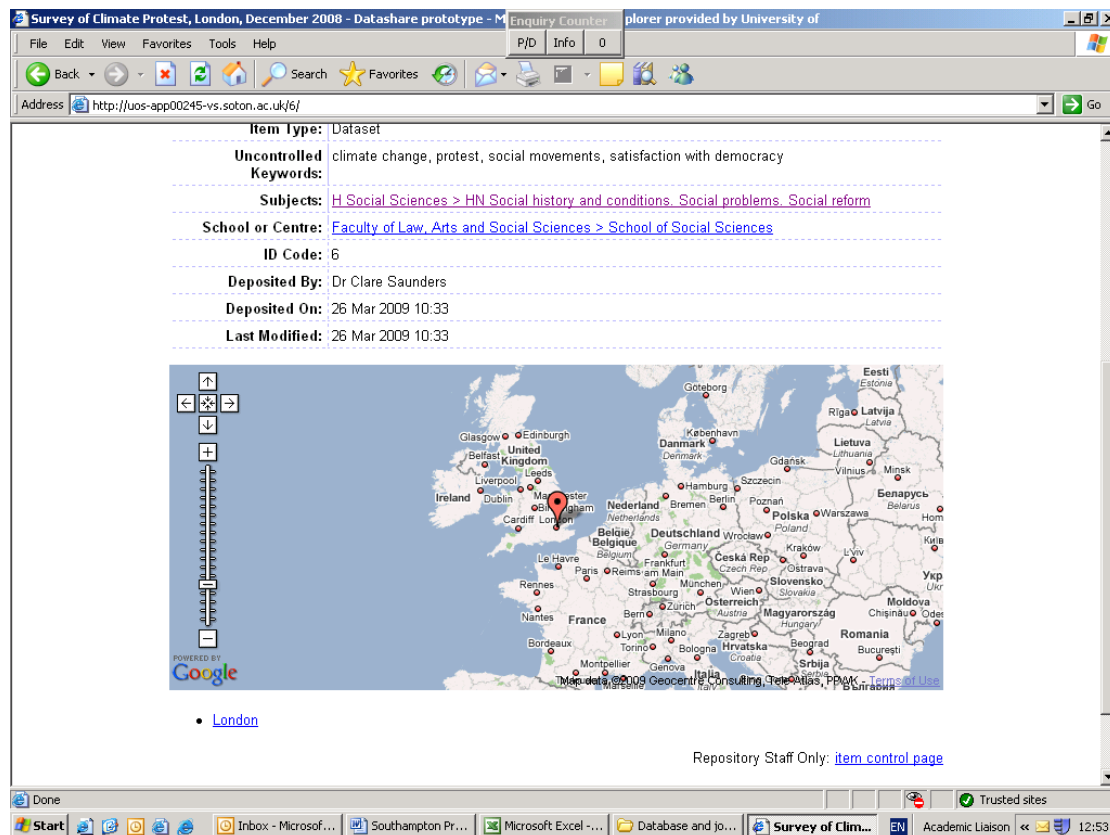


Figure 2: Geonames Google Maps display in DataShare Prototype

## Assessing Impact

Possibly the most significant impact of DataShare at Southampton is its influence on institutional policy and strategy through communication with the University's Data Preservation Group. DataShare has directly informed the decision to bid for Roberts' money to fund training for post-graduate research students in areas including data management. The experience gained through the Data Survey will also inform a series of data life cycle interviews which are due to take place with researchers outside of social sciences. Furthermore, the Data Survey established a model for partnership working between the Library and the research community that will be taken forward in future initiatives.

DataShare has also contributed a substantial development to Southampton's institutional repository, ePrints Soton. When the upgrade is complete, ePrints Soton will be in a position to accept research datasets. It will also contain at least four exemplar datasets that will be used to demonstrate and promote the service.

Beyond the University of Southampton, development of the DataShare Prototype has resulted in the addition of Geonames functionality to ePrints software that has the potential to benefit ePrints users throughout the world.

Although it is too early to assess the full impact of the Data Survey, it is felt that it has already raised the profile of the Library in terms of data support. Since the survey, the Library has received two data queries from researchers in the Division of Social Statistics, when previously data queries came almost exclusively from taught students in Economics. In turn, the 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.

**London School of Economics DataShare Report  
(March, 2009)**

The LSE was an original member of the DataShare project, but was regrettably forced to withdraw due to staffing and recruitment problems. However the university is still involved in investigating repository services.

LSE Research Online was set up in 2005 and uses EPrints software. This is a central database where research from all LSE departments can be searched. As well as standard searches by year, author, publication, it is also possible to search by department or research centre. Additionally the repository is linked to LSE Experts allowing a search of publications by area of expertise.

LSE Research Online currently includes over 16,000 publications of which almost 3,000 are full text records. During the next 12 months the team will be investigating the possibility of including other objects such as statistical datasets, images etc. However it is possible to link publications in LSE Research Online to datasets held in Dataverse.

The LSE is also involved with Economists Online, a service developed by the NEEO project which is the flagship project of the Nereus consortium. NEEO is funded by the EU to ‘address the lack of integration of academic output amongst premier economics institutions’. Datasets accessible via Economists Online are stored in Dataverse, and the NEEO project is in the process of investigating the legal, ethical and organisational challenges faced by researchers and repositories wishing to make statistical datasets available. As of the 31<sup>st</sup> March 2009, 897 economists from 19 institutions in 9 countries (including Columbia in the US) were contributing publications to Economists Online.

LSE Research Online  
Eprints  
NEEO / Economists Online

<http://eprints.lse.ac.uk/>  
<http://www.eprints.org/>  
<http://www.neeoproject.eu/>  
<http://www.economistsonline.org>  
<http://thedata.org/>

Dataverse