# Digital Repository Services for Managing Research Data: What Do Oxford Researchers Need?

*Luis Martinez-Uribe**

**Abstract**
UK researchers are facing the challenges of having to comply with funder requirements to submit data management plans and make their data available. Academic institutions have the responsibility to support their researchers to fulfill their contractual requirements with funding agencies. Increasingly, repository services are dealing with the management of research output. Managing research data to ensure digital materials adhere to the right standards and are securely stored, shared and preserved can be complex and resource intensive. Understanding how researchers work is the key to designing university repository services to manage research data. This article describes the University of Oxford's federation of digital repositories and introduces the Scoping Digital Repository Services for Research Data Management project to present the findings from the requirements gathering exercise carried out to understand Oxford researchers' practices and needs. .

**Introduction**
The proliferation of gadgets that deliver information ming In the current climate of large scale international projects around research data such as the Australian National Data Service, the US National Science Foundation DataNet and the UK Research Data Service, the management of research data is a topic that attracts interest from policy makers worldwide because of the importance of data in the age of the knowledge economy (PMSEIC 2006). There are many efforts to foster standards for data description and sharing as well as to establish national and international federations of digital repositories to deal with the management and curation of these digital resources. UK Universities and their researchers are facing the challenges themselves. Researchers in all disciplines are increasingly being asked by funding bodies to not only make their data available but also to submit data management and data sharing plans with their funding applications. These plans should describe what datasets will be created that are worth keeping, what standards will be used, how they will be made available and who will be responsible for their long-term preservation (Weaver 2007). Managing and curating research data poses many challenges because of their heterogeneity and a lack of standards as well as many unresolved ethical issues (Carusi & Jirotka 2008). Some of the benefits of the active management and curation of research data include the possibility to replicate research results, avoiding expensive data collection by promoting data reuse and protecting the contributor's sensitive information (Schroeder & Axelsson 2007).

When planning how best to manage and curate research data, efforts are sometimes mainly focused on understanding the data themselves, the different types, their volume and specific technical or legal issues surrounding them. Nonetheless engaging with the producers and users of those datasets is at times forgotten. Understanding researchers' needs and workflows will help to comprehend how they work with data, why they create them in the first place and their reasons for managing these resources the way they do. This approach will also assist to identify services to support them throughout their research process so that they can fulfill the requirements from funders.

This paper describes briefly Oxford's federation of digital repositories and then introduces the Scoping Digital Repository Services for Research Data Management project and the findings of the requirement gathering exercise carried out between May and June 2008 as part of the project.

**Background: A Federation of Digital Repositories in a Collegiate University**
The University of Oxford, the oldest university in the English speaking world, has a complex structure with divisional departments, institutes, independent colleges and more than a hundred libraries. A highly devolved institution which has on many occasions been compared to a microcosm of the entire UK Higher Education system; this organizational arrangement is mirrored with a devolved computing structure (Jeffreys 2008). The main central ICT service provider is Oxford University Computing Services (OUCS) that operates the primary computing infrastructure such as core networks, back up servers and core support services. Another central service provider providing ICT for its users is the Oxford University Library Services (OULS). These centralized services are complemented with local ones embedded in departments, institutes and colleges with their own ICT infrastructure and support teams.

Acting as an overall umbrella is the Office of the Director of IT (ODIT) providing strategic direction for IT at the University.

In terms of digital repositories, Oxford can be seen as a federation with the Oxford Digital Repositories Steering Group providing a coordinating role for the development of repository services, see figure 1. The Oxford Research Archive (ORA) is the digital repository infrastructure for OULS providing permanent and secure online archive of research output materials produced by members of Oxford. Its content includes journal articles, conference papers, working papers, theses and other grey literature. However, there are a number of other digital collections and repository activities like the data resources from the Oxford e-Research Centre (OeRC), the Google materials, the Oxford Digital Library Collections and others. This wealth of digital resources, the "increasing need to manage curation and access to research data" (Fraser 2005) and the lack of interaction between activities led to the creation of the Digital Repositories Research Coordinator post and the start of the Scoping Digital Repository Services for Research Data Management project.

## The Scoping Digital Repository Services for Research Data Management Project

The Scoping Digital Repository Services for Research Data Management project is a joint effort between ODIT, OUCS, OULS, OeRC and reports to the Oxford Digital Repositories Steering Group. The project scopes the requirements, including infrastructure and interoperability, for repository services to manage and curate research data generated by Oxford researchers.

Before describing the project any further it is important to clarify the scope of what is meant here by research data, data management and repository services.

Research data are a heterogeneous type of research output which can take many forms (text, numbers, audio, images, moving images, etc) and might be created for different purposes during the research process. The National Science Foundation provides a useful categorization based on the origins of the data: observational, computational or experimental (2005). Observational data are historical records such as opinion polls or precipitation measurements. These data cannot be recollected, thus the importance of preserving it indefinitely. On the other hand,

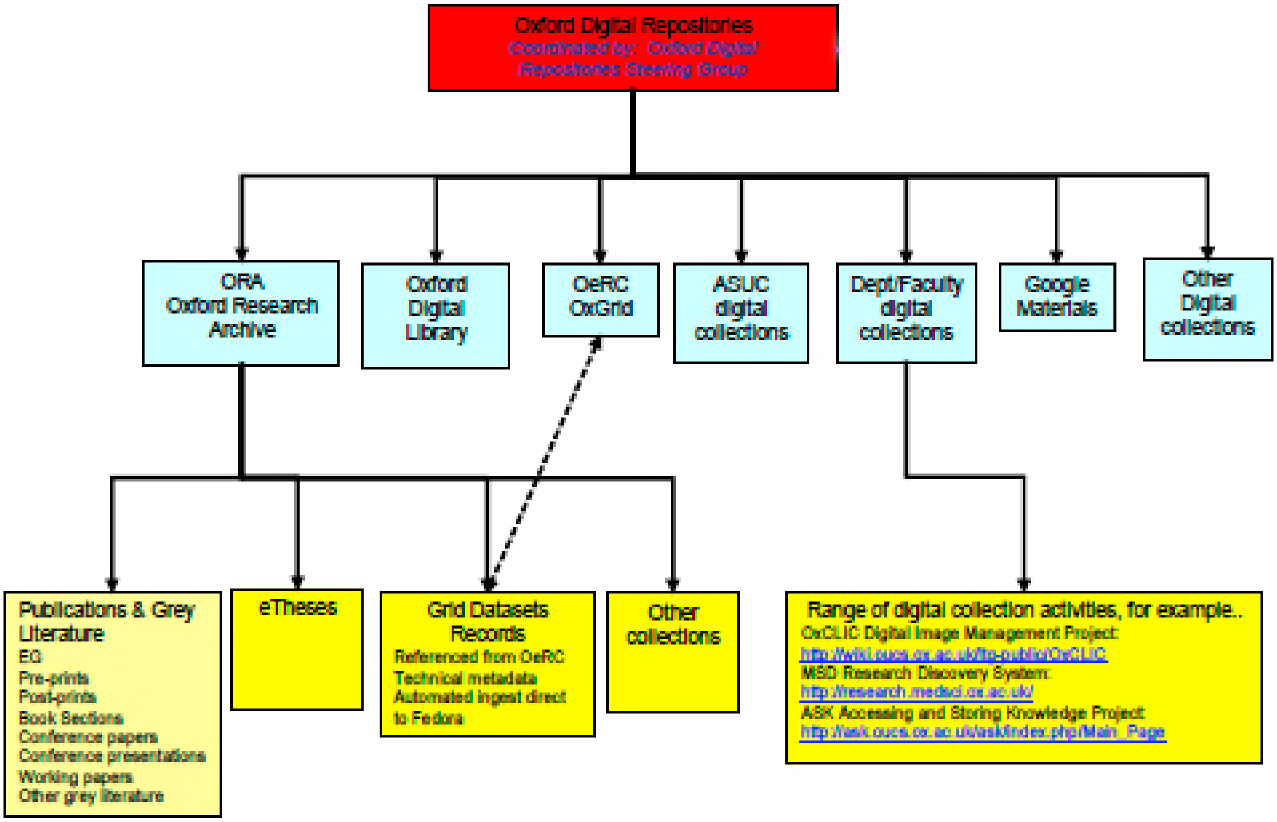

Oxford Digital Repositories
Coordinated by: Oxford Digital Repositories Steering Group
ORA Oxford Research Archive
Oxford Digital Library
OeRC OxGrid
ASUC digital collections
Dept/Faculty digital collections
Google Materials
Other Digital collections
Publications & Grey Literature
EG
Pre-prints
Post-prints
Book Sections
Conference papers
Conference presentations
Working papers
Other grey literature
eTheses
Grid Datasets Records
Referenced from OeRC
Technical metadata
Automated ingest direct to Fedora
Other collections
Range of digital collection activities, for example..
OxCLIC Digital Image Management Project:
http://wiki.oucs.ox.ac.uk/ltg-public/OxCLIC
MSD Research Discovery System:
http://research.medsci.ox.ac.uk/
ASK Accessing and Storing Knowledge Project:
http://ask.oucs.ox.ac.uk/ask/index.php/Main_Page

Figure 1. Oxford Digital Repositories Structure (Jeffreys & Fraser 2007)

computational data resulting from computing simulations can be reproduced. Preserving the input files that allow replicating the simulation is more important than preserving the raw data obtained through the simulation. Data produced through experiments poses other challenges. In many cases although the experiment could be reproduced this may be too costly. The Research Information Network (2008) adds two more categories in their data typology: derived and canonical data. The former refers to data resulting from some form of processing to primary data while the latter refers to those reference datasets such as the gene sequence.

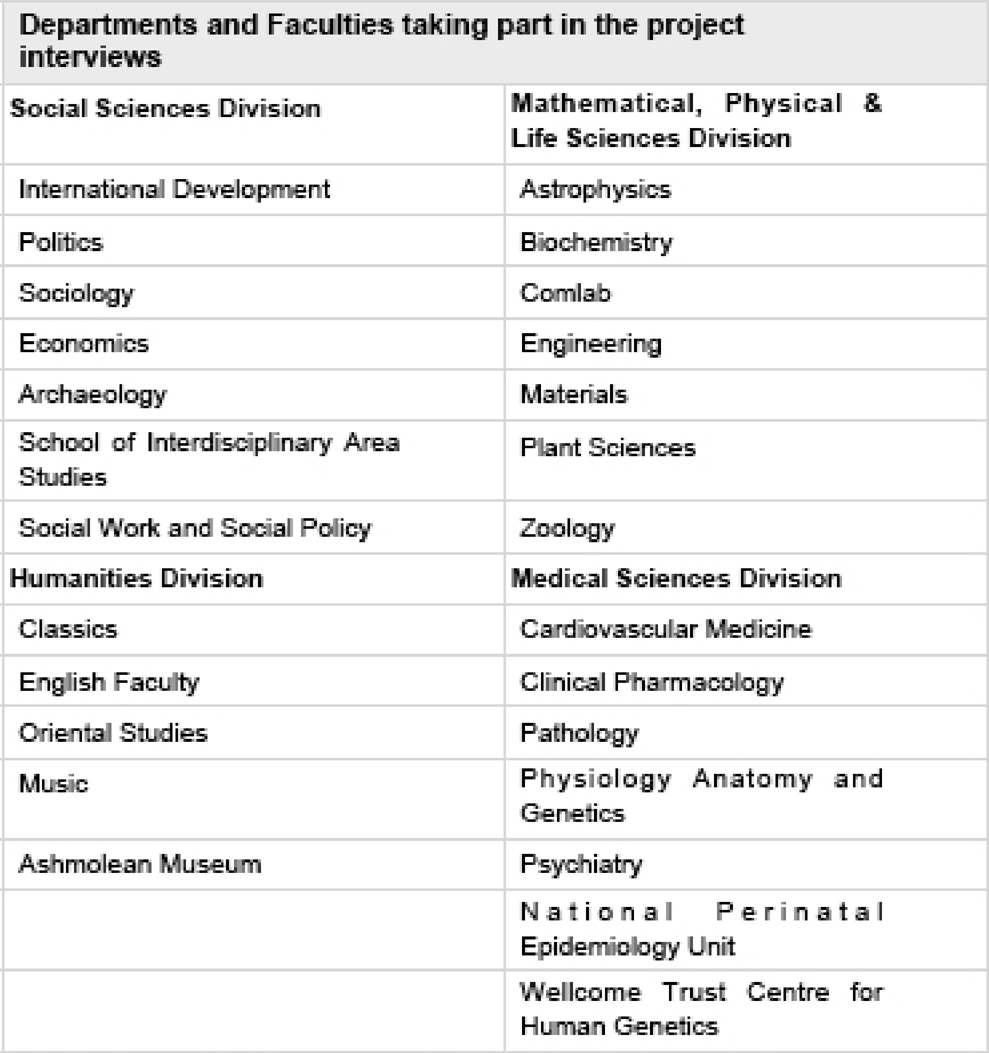
Data management is a vast discipline and includes activities such as database design, data compliance and data modeling, and many others. It takes from disciplines like information and knowledge management that focus on looking after these assets from the moment of creation and dissemination through the organization. These activities include storage, retrieval, use, access, preservation or disposal (Macevi & Wilson 2005; Alavi & Leidner 2001). The US Department of Defense defines this as "understanding current and future data needs" (Parker 2000) and as pointed out by a study from Virginia Commonwealth University (Aiken et al 2007) the management of data needs to be seen as a means to an end. In this project one of the main drivers for improving the current infrastructure for data management is the pressing requirement from funding bodies to make data available and provide data management and data sharing plans with funding applications.

The term "repository services", in its widest sense, refers not only to a technical infrastructure that allows storage, access, description, dissemination and preservation of digital objects but also the support services to assist researchers with technical and legal issues and policies for the creation, deposition and sharing of digital research outputs.

**Requirements Capture: Methodology and Participation**

One of the main objectives of the project was to document data management practices of Oxford researchers as well as to capture their requirements for services to help them manage their data more effectively. In order to do this, thirty-seven face-to-face interviews were conducted between May and June with researchers from twenty-six departments and faculties from Oxford, see table 1. This positive response to the interview calls provided a good cross-section: 58% of which were on the ground researchers; 28% Heads of departments/faculties or research teams; and the remaining 14% a mixture of Data Managers, IT Officers and Administrators embedded in departments/faculties and research teams.

The interview framework was largely based on the methodology employed by other Oxford projects (the e-Infrastructure Use Cases,  Building a Virtual Research Environment for Humanities  and the Integrative Biology



| **Departments and Faculties taking part in the project interviews** | |
|---|---|
| **Social Sciences Division** | **Mathematical, Physical & Life Sciences Division** |
| International Development | Astrophysics |
| Politics | Biochemistry |
| Sociology | Comlab |
| Economics | Engineering |
| Archaeology | Materials |
| School of Interdisciplinary Area Studies | Plant Sciences |
| Social Work and Social Policy | Zoology |
| **Humanities Division** | **Medical Sciences Division** |
| Classics | Cardiovascular Medicine |
| English Faculty | Clinical Pharmacology |
| Oriental Studies | Pathology |
| Music | Physiology Anatomy and Genetics |
| Ashmolean Museum | Psychiatry |
| | National Perinatal Epidemiology Unit |
| | Wellcome Trust Centre for Human Genetics |

*Table 1. Departments and Faculties taking part in the project interviews*

Virtual Research Environment ) with some adjustments to fit with the requirements of this study. The interviews were semi-structured to engage in conversations with researchers and delve into participant reasons for doing things they way they do. A generic research life cycle model was referenced to structure the conversation. The research life cycle started with the funding application, moved into data collection and data processing, and finished with data publication. Several approaches were used to identify interview candidates: the first choice of interviewees was guided by suggestions from members of OUCS, OULS and OeRC and then a call for interviewees circulated amongst research facilitators. For every researcher interviewed, a snowball sampling approach was used to identify further candidates.

In addition to this an event, the Research Data Management Workshop, was organized to complement interview findings. This event brought 46 attendees throughout the day from 24 departments, research centres and colleges. Both the interviews and the workshop also formed the basis of the Oxford case study for the UK Research Data Service feasibility study, a joint project between Research Libraries UK and the Russell Group IT Directors Group, aiming to assess feasibility and cost of developing a national shared data service for research data generated in UK Higher Education Institutions.

**Findings: From funding application to data publication**
Findings from the interviews and the workshop revealed that the management of research data occurs with varying degrees of maturity across Oxford University. There are some departments that have been dealing with very sensitive data for several years and they have policies and procedures in place that support a robust technical infrastructure. On the other hand, many other units in the University tend to work on a more ad-hoc basis and data management relies on the individual researcher's skills. Overall, Oxford researchers felt there were potential services to help them manage their data more effectively.

At the funding application stage, researchers tend not to plan the management of the data at the outset of their research project in detail. As one of the researchers interviewed stated "…you are not interested in this because you are interested in the science; the technical issues come up later." With the wide variety of funding sources available, they find it confusing to understand what the different requirements are for making their data available and the retention period.

The types of data collected were, as expected at the beginning of the project, many and very diverse. There exist a significant variety of forms and formats, some of them proprietary. In addition, there are a wide range of sizes, from few megabytes to several petabytes. Some of the data were highly sensitive, and strict ethical and security protocols needed to be followed to collect these data. The long-term usefulness of the data also varied enormously: in some fast moving disciplines the data would be relevant up to five years before better data could be produced, while in other cases it is impossible to reproduce the results, and the life-span is indefinite.

Once the data are collected they are mostly stored on personal computers or departmental servers with a variety of security and back up procedures; although there are some horror stories about shelves full of highly valuable data stored on CDs and DVDs. Metadata accompanying the data tends to be minimal and data are organized in hierarchical folder structures with file names that make sense to the researcher. These data are then shared in informal ways and this happens mostly by email or portable media. Problems arise when the size of the data increases and then the storage and sharing becomes difficult.

Very few of the researchers interviewed had deposited any data in domain specific data archives such as the Natural Environment Research Council data centres  or the UK Data Archive . Nonetheless, many of them were publishing data on their departmental websites. Data ownership is seen as a major issue. There have been cases where data have been generated from human subjects as part of collaborative research projects between many institutions in different countries and with several funding bodies. In cases like these and others, researchers struggle to understand who owns the data produced in their research projects. In terms of sharing their data, although researchers tend to feel very attached to their data, they believe that if their research is publicly funded, then their data should be made publicly available.

The top three requirements expressed by Oxford researchers for services to help them manage their data more effectively were:

- A secure and user-friendly solution that allows storage of large volume of data and sharing of these in a controlled fashion, allowing fine grained access control mechanisms. \

- A sustainable infrastructure that allows publication and long-term preservation of research data for those disciplines not currently served by domain specific services such as the UK Data Archive, NERC Data Centres, European Bioinformatics Institute and others.

- Advice on practical issues related to managing data across its life cycle. This help would range from assistance in producing a data management/sharing plan; advice on best formats for data creation and options for storing and sharing data securely; to guidance on publishing and preserving these research data.

## Conclusion and Next Steps

As pointed out by Lyon (2007), in order to manage and curate research data it is crucial that the different communities of researchers, librarians and computing services work together. Nonetheless, in order to deploy an effective and usable infrastructure for managing research data it is key to understand the producers of the data themselves, their workflows and needs. The interviews and workshop have provided enough evidence about current data management practices at Oxford and researchers' requirements for services to help with these. The findings will be complemented with another workshop and a consultation with service providers in Oxford to assist in producing a set of recommendations to improve and coordinate the provision of digital repository services for research data at Oxford[16.]

* Contact: Luis Martinez-Uribe, Digital Repositories Research Co-ordinator, Oxford e-Research Centre, University of Oxford. E-mail: luis.martinez-uribe@oerc.ox.ac.uk

## References

Aiken, P., D. Allen, B. Parker & A. Mattia (2007 April) Measuring Data Management Practice Maturity: A Community's Self-Assessment. Computer. Retrieved March 15, 2008

Alavi, M. & D. E. Leidner (2001) Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues. MIS Quarterly, 25, 107-136

Carusi, Annamaria & Jirotka, Marina.(2008). From data archive to ethical labyrinth. Qualitative Research, Forthcoming

Jeffreys, Paul. (2008). Oxford's Computing Model : Delivering Computing Services to the University. Retrieved July 15,2008 from: http://www.ict.ox.ac.uk/odit/ITcoordination/ODITOxford%27sComputingModel.pdf

Jeffreys, Paul & Fraser, Mike (2007). Oxford Digital Repositories Steering Group Meeting Agenda - Job description for Digital Repositories Research Co-ordinator – Annex 1: Oxford Digital Repositories Structure. Retrieved July 15,2008 from: http://www.ict.ox.ac.uk/repositories/meetings/ODRSG_agenda-papers_4May07_2.pdf

Fraser, Mike. (2005). Towards a Research Repository for Oxford University.Retrieved March 10,2008 from: http://ora.ouls.ox.ac.uk/objects/uuid:5fa206c9-c400-403c-ab69-174ce8604a7a

Lyon, Liz. (2007). Dealing with Data: Roles, Rights, Responsabilities and Relationships – Consultancy Report. Retrieved May 17,2008 from: http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf

Macevi, E. & T. Wilson. (2005). Introducing information management: an information research reader. London, UK : Facet Publishing

National Science Foundation. (2005). Long-Lived Digital Data Collections; Enabling Research and Education in the 21st Century. Retrieved March 15, 2008 from http://nsf.gov/pubs/2005/nsb0540/

Parker, B. (2000). Enterprise Data Management Process Maturity. In Data Management Handbook. Auerbach Publications.

PMSEIC (Prime Minister's Science, Engineering and Innovation Council) Working Group on Data for Science. (2006). From Data to Wisdom: Pathways to Succesful Data Management for Australian Science. Retrieved March 12, 2008 from: http://www.dest.gov.au/sectors/science_innovation/publications_resources/profiles/presentation_data_for_science.htm

Research Information Network.(2008). Stewardship of Digital Research Data : A Framework of Principles and Guidelines. Retrieved July 15,2008 from http://www.rin.ac.uk/data-principles

[proceedings] Schroeder, Ralph & Axelsson, Ann-Sofie. (2007). Making it Open and Keeping it Safe: e-Enabled Data Sharing in Sweden and Related Issues. E-Social Science 2007 Ann Arbor, Michigan US. Retrieved July 2008 from: http://ess.si.umich.edu/papers/paper139.pdf

Weaver,Belinda. (2007). Constructing a Research Project Data Management Plan. Creating a data management strategy for new research projects Workshop, University of Queensland, Australia. Retrieved on July 15, 2008 from: http://www.library.uq.edu.au/escholarship/orca.html

## Footnotes

1. http://www.ands.org.au/

2. Sustainable Digital Data Preservation and Access Network Partners (DataNet) http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503141&org=OCI&from=home

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

4. http://ora.ouls.ox.ac.uk/

5. http://www.oerc.ox.ac.uk/resources

6. http://www.bodley.ox.ac.uk/google/

7. http://www.odl.ox.ac.uk/

8. An up to date list of activities can be found at: http://www.ict.ox.ac.uk/repositories/index.xml.ID=body.1_div.6

9. http://www.ict.ox.ac.uk/odit/projects/digitalrepository/

10. www.eius.ac.uk/

11. http://bvreh.humanities.ox.ac.uk/

12. http://www.vre.ox.ac.uk/ibvre/

13. digitalrepository/Workshops.xml

14. http://www.nerc.ac.uk/research/sites/data/

15. http://www.data-archive.ac.uk

16. A detailed report of the findings from the interviews and the workshop, next steps after the interviews and other outputs from the project can be found at: http://www.ict.ox.ac.uk/odit/projects/digitalrepository/