FINDINGS OF THE SCOPING STUDY INTERVIEWS AND THE RESEARCH DATA MANAGEMENT WORKSHOP

SCOPING DIGITAL REPOSITORY SERVICES FOR RESEARCH DATA MANAGEMENT

A Project of the Office of the Director of IT

www.ict.ox.ac.uk/odit/projects/digitalrepository/

Author	Luis Martinez-Uribe (luis.martinez-uribe@oerc.ox.ac.uk) Digital Repositories Research Co-ordinator
Affiliation	Oxford e-Research Centre
Version	1
Date created	27/5/08 9:52
Date last modified	Last saved by Luis Martinez-Uribe on 25/7/08 10:29
Distributed to	Paul Jeffreys, Richard Ovenden, Mike Fraser

Contents

EXECUTIVE SUMMARY

The project Scoping Digital Repository Services for Research Data Management is a joint effort in between the Office of the Director of IT, the Oxford University Computing Services, the Oxford University Library Services and the Oxford e-Research Centre. The project reports to the Oxford Digital Repositories Steering Group and aims to scope requirements, including underlying infrastructure and interoperability, for services to manage research data generated by Oxford researchers. One of the main activities of the project is the scoping study interviews, a requirement gathering exercise to learn about current data management practices and identify top requirements from researchers for services to help them manage their data more effectively. To complement the findings of the interviews, as well as to raise awareness and encourage discussion, a workshop was organized in June to hear about examples of good and interesting practice with respect to the use of digital repository services at various points in the research life cycle and from the perspective of different disciplines. Findings from the project and the workshop also form the basis for the Oxford case study for the UK Research Data Service (UKRDS) feasibility study for a shared national data service. This document presents the findings of the interviews undertaken during May and June 2008 and the complementary workshop.

Findings and Top Requirements for Services

A total of 37 researchers took part on the scoping study interviews and 46 people attended the workshop in June. This good response from researchers reveals the interest on research data management and it allowed to document current practices and to capture requirements for services across disciplines in Oxford.

The management of research data in the University of Oxford is exercised to variable degrees of maturity across the institution. There are departments and individuals with an extensive experience in handling the data they collect and big projects with a focus on data activities which produce, document and share data to a very high standard. On the other hand, there are many other departments and small-scale projects in which the data management depends entirely on individual researchers skills and this is sometimes driven by individual short-term convenience.

Overall, the vast majority of researchers interviewed thought that there are potential services that could help them manage their data more effectively.

The top requirements from Oxford researchers for services to help with their data management activities gathered from the interviews and the workshop are:

Advice on practical issues related to managing data across their 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.

A secure and user-friendly solution that allows storage of large volume of data and sharing of these in a controlled fashion way allowing fine grain 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.

Funding that could help address some of the departmental challenges to manage the research data that are being produced.

INTRODUCTION

Researchers produce vast amounts of digital material during the course of their research activities. Some of these materials can be very costly to produce, are crucial to replicate research results and extremely valuable to other researchers who may be able to use them in novel ways. Research data are part of the previous resources and form a heterogeneous and complex-to-manage set of digital materials. Funding agencies have realised the importance of keeping them safe and are increasingly requiring research projects to ensure that any data produced, that may be useful to others, are of quality as well as shared and preserved. In order for the data to be of quality and useful to others, they need to be properly managed from the moment of creation and going through the organization. Research data are valuable assets for the Departments and Institutes of Oxford University and it is important that services are in place to facilitate the management of these resources.

This document contains the findings from the scoping study interviews conducted in Oxford University during May and June 2008 and a complementary workshop for Oxford researchers on data management held on the 13th of June. The interviews and the workshop are part of the activities of the Scoping Digital Repository Services for Research Data Management project, a joint effort between the Office of the Director of IT, the Oxford University Computing Services, the Oxford University Library Services and the Oxford e-Research Centre. The scoping study is aimed at scoping the requirements including underlying infrastructure and interoperability for digital repositories services to store, disseminate and preserve research data generated at Oxford. In addition to this, the findings from the interviews and workshop form the basis for the Oxford case study for the UK Research Data Service (UKRDS) feasibility study[1]. The UKRDS is a joint project between Research Libraries UK (RLUK) and the Russell Group IT Directors Group (RUGIT) aiming to assess feasibility and cost of developing a national shared data service for research data generated in UK Higher Education Institutions.

The aim of this paper is to inform the Oxford Digital Repositories Steering Group, to which the project reports, other relevant stakeholders in Oxford and the UKRDS Steering Committee on current research data management practices within Oxford and present researchers' top requirements for services to assist them in this task.

The document is organised as follows: explanation of the methodology used for the interviews; presentation of the findings using six major categories: Funding, Data Collection, Processing of Data, Publishing, Support and Top Requirements. The complementary findings of the workshop follow together with selected case studies derived from presentations made by Oxford researchers. The final section of the report outlines some conclusions and the recommended next steps for the project.

[1] http://www.ukrds.ac.uk/

1. INTERVIEW METHODOLOGY

In this section the methodology used for the scoping study interviews is described to explain how the interview candidates were identified, the interview process that followed and how the analysis was carried out. The interview framework used was largely based on the methodology employed by the e-Infrastructure Use Cases (eIUS) project[2], Building a Virtual Research Environment for Humanities (BVREH) project[3] and the Integrative Biology Virtual Research Environment (IBVRE)[4] project with some adjustments to fit with the requirements of this study.

1.1 Scope

The scoping study interviews aimed to capture and document data management practices to identify best practice and service gaps in the current infrastructure. The focus was primarily on Oxford researchers and the data they produce during their research activities whilst recognizing that the researchers themselves traverse institutional and national boundaries and that data may be different materials in the different disciplines. All research disciplines were within scope as well as different roles, ranging from Heads of Department to Research Students, to achieve a good cross-section.

1.2 Identification of Interviewees

In order to identify suitable candidates to interview a combination of several approaches was adopted. Choice of candidates was originally guided by suggestions from members of the Oxford e-Research Centre, Oxford University Library Services and Oxford University Computing Services. In addition to this, a paper introducing the project was presented at the Medical Science Division Research Committee meeting in April and further candidates were put forward for interview. The Research Services Office helped to reach a wider audience by contacting their research facilitator network to circulate the project briefing and invite researchers to participate in the scoping study interviews. A call for interview volunteers was also published on their news bulletin. This call was also posted in the project blog with a “space for comments” created to encourage researchers who could not take part in the interviews to specify their data management requirements. Academic members of the Digital Repositories Steering Group as well as those consulted for the UK Research Data Service were contacted to suggest further candidates. Finally for each researcher identified, the *friend of a friend* approach or snowball sampling was used to populate the list of interview candidates.

1.3 Organization of Interviews

The interviewees identified were emailed individually providing them with a project brief and asking them to take part on the scoping study. When a suitable interviewee was identified, the interview questions were circulated and time and place was arranged for the interview to take place. In order to improve the interview framework, a few pilot interviews were conducted where interviewees were asked to provide feedback on the interview process.

1.4 The Interview Process

Interviews lasted no longer than one hour and they required some preparation to learn more about the interviewee’s specific area of research. The interview itself, see appendix1, were semi-structured and started with a brief introduction to the project and a reminder of the nature of the interview as well as the intention to tape it with a

[2] www.eius.ac.uk/
[3] http://bvreh.humanities.ox.ac.uk/
[4] http://www.vre.ox.ac.uk/ibvre/

recorder, with permission, and take notes. During the interview, it was attempted to understand the information produced during the research process and the approach taken to manage this to comprehend whether this is subject to individual convenience or departmental procedures or guidelines.

1.5 Data Analysis

After each interview a two pages summary was produced. The analysis of notes and recordings of each interview was then integrated with findings of other projects run in Oxford like eIUS, BVREH or IBVRE to produce the content of the next section.

2. FINDINGS

2.1 Distribution of Interview Respondents

A total of 37 researchers participated in the scoping study interviews with representatives from the departments and faculties shown below in table 1. As shown in figure 1, 58 % of respondents were *"on the ground"* researchers (including DPhil students), 28 % Heads of departments or research groups and 14 % technical staff, Data Managers or administrators embedded in research teams and departments.

Departments and Faculties

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 taking part in the scoping study interviews

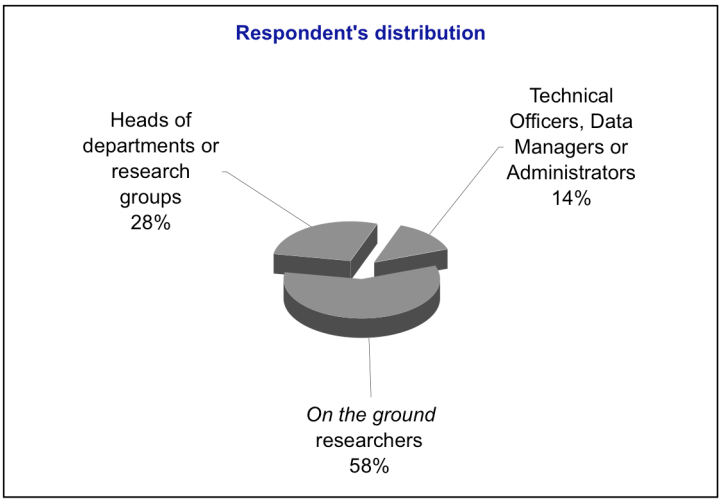

Figure 1. Interview respondents distribution

2.2 Interview Findings

Thanks to the good response from Oxford researchers to the interview call, it has been possible to explore through this study how the management of data takes place and researchers' needs for services to help them with it. There was a great interest on the topic across disciplines although in certain domains, like Medical Sciences, there seems to be a real and pressing need to improve current practices *"issues around medical data are highly topical, controversial and cause a lot of anxiety".* There was also a general recognition that things could be done better and the need for someone to take the lead *"…we are not very smart on data management but no one wants to take responsibility".*

The management of research data in the University of Oxford is exercised to variable degrees of maturity across the institution. There are departments and individuals with an extensive experience in handling the data they collect and big projects with a focus on data activities which produce, document and share data to a very high standard. On the other hand, there are many other departments and small-scale projects in which the data management depends entirely on individual researchers skills and this is sometimes driven by individual short-term convenience.

Overall, the vast majority of researchers interviewed thought that there are potential services that could help them manage their data more effectively.

The interviews were structured around a generic *research life cycle* model to add some form of structure to the conversation held with participants. The life cycle used starts with the funding application followed by the collection and processing of data to finish with the publication. As originally anticipated, a real research life cycle is by no means as neat as the proposed here, researchers were not necessarily involved in all the phases of the proposed cycle and provided answers to the questions most relevant to their roles. The questions using the life cycle approach were followed by other questions on support and top requirements for services. These headings are used next to present the findings.

2.2.1 Funding

This part of the interview was included in order to better understand how researchers think about data at the early stage of applying for funding and how well they are aware of their responsibilities with funders in terms of data sharing and archiving.

- At the funding stage researchers tend to not plan in detail the data management that will need to be carried out but further down the line during their researcher activities they face problems and challenges to manage the data they produced:

  "…when putting your grant together you are not interested in this because you are interested in the science, the technical issues come up later"

- Researchers and research groups get funding from a variety of sources with variable requirements in terms of data management and data sharing plans. Increasingly the projects are multi-disciplinary, across departments and institutions, funded by several funding agencies and thus the difficulty for researchers of knowing what are their obligations and responsibilities towards data collected. Moreover, in cases like the

previous and others researchers struggle to understand who owns the data produced in their research projects.

- Mostly researchers have not had to provide data management plans so far as part of the funding application. The ones that have made data management plans, mostly when the data is highly sensitive, have either got support from their technical staff or produced generic plans describing the types of data that will be created, how they will be stored, backed-up and made available to others.

- Mostly researchers are aware that data will need to be made available at the end of the research project but also know that this is not policed or enforced. Researchers in Humanities are concerned about the cessation of the Arts and Humanities Data Service (AHDS) whilst the Arts and Humanities Research Council (AHRC) still requires them to make their data available.

- The situation with clinical trial units is rather pressing at the moment. The UK Clinical Research Collaboration (UKCRC) has been set up to re-engineer the environment in which clinical research is conducted in the UK and are now registering units with high quality trials expertise. They have to demonstrate a number of key competencies and one of them is data management. Becoming a member brings benefits like access to the infrastructure of the UKCRC, being able to collaborate with others groups and share information with them but above all, getting funding to run trials.

2.2.2 Data Collection

This section was originally designed to learn about the different ways in which data are collected and captured, the different types of formats and sizes as well as the apparent usefulness of these resources to others.

- The collection of data occurs in supercomputers generating huge simulations; in laboratories with sophisticated instruments such as microscopes or scanners producing large scale images, from field work in Social Science interviewing human subjects or archaeological excavations generating pictures, annotations and maps; from health professionals filling paper questionnaires when examining their patients; from research with manuscripts undertaken in libraries and museums worldwide; etc.

- There is significant work done by researchers collecting data from published articles and books. This data collection involves manually inputting several datasets included in the publications into spreadsheets to then analyse them.

- The collection of data can be highly expensive. A particular project claimed to spend around £500 K collecting data that are unique and of high value to other researchers.

- The reproducibility of the data varies too, some data is not possible to be collected again or it may be too costly to do so. In other cases, like simulations, the algorithm and the input files are more important than the actual data generated by them.

- A variety of types of data collected. These can be nicely grouped using the Research Information Network data typology[5]: data from observations that capture some measurement of a phenomenon in a location at a given point in time; generated by computer simulations to test models; data from experiments in scientific laboratories using sophisticated equipment like microscopes or scanners; derived data as a resulting product of manipulation and processing of primary data and canonical or reference data like those relating to gene sequences or literary texts.

- Research data produced by Oxford researchers comes in a variety of formats that include text (flat text files, MS Word, Word Perfect), numerical (MS Excel, MS Access, MySQL, SPSS, STATA), video, audio, images (jpeg, tiff). Some of these data come in proprietary formats that can only be used with specific software products. There are a wide range of sizes, from a few megabytes in disciplines like Humanities to terabytes generated by simulations in Medical Sciences and MPLS.

- The usefulness of the data to other researcher varies depending on the discipline. Some of the data created in fast moving scientific disciplines may be of value for the next five years whilst data created in areas such as Social Sciences or Art and Humanities may be useful indefinitely. Not always the raw data are useful to other researchers.

- Secondary data is used by most of the researchers interviewed. These data are found through either well-known discipline specific data repositories or obtained in an informal basis through networks of contacts from conferences or similar events.

2.2.3 Processing of Data

In this portion of the interview the aim was to understand researcher's workflows to store data securely, access them to manipulate them and sharing these data with collaborators.

- Data are commonly stored on personal computers or departmental servers. The data produced as part of computing simulations mostly in the divisions of MPLS and Medical Sciences is enormous in size and researchers struggle to find secure alternatives. As one of the interviewees stated: *"there is nowhere in Oxford where people running simulations can store their data"*

- The data tends to be so precious to researchers that almost all of them have back-up strategies and mostly use OUCS services for this purpose. Nonetheless, some research groups use CDs and DVDs to back up and archive their data.

  "… he is storing all of his data for the last fifteen years on DVDs and he has teens of terabytes. They didn't think about it at the outset and they are not technical. There are horror stories around the university about how the data is stored"

- Annotation of the data to record its provenance and content takes place mostly by including the information within the data, using hierarchical folder structure with file names and *readme* files in some cases. This

[5] Research Information Network (2008) "Stewardship of Digital Research Data : A Framework of Principles and Guidelines"

information helps the researcher using the data while working with them but it would be hard for others to make sense of it. After some time even the researcher that created the data in the first place would find difficult to reuse the data, as one of them stated: *"in five years time I wouldn't know what was going on".*

- In some big projects where data and data curation have played an important role, high quality metadata has been created to describe the data produced. Nonetheless, very few researchers are aware of existing standards to describe their data.

- In research projects or units that generate sensitive data policies and procedures are in place to store and access these data securely. Research groups that work with sensitive data have experience in anonymizing these.

- Most researchers share the data they work with and this again happens in many ways. Some will use email if the files are not too big, they will upload it into a website and share the URL, copy it into portable media to then mail them, etc. The problem arises when the size of the data does not allow any of these methods or when the data is so sensitive that ethics approval is needed. During the interviews a situation come up where researchers had to copy the data from a storage device to another and physically take it to the collaborators.

- Some of the tools for analyzing the data include R, SPSS, STATA, MATLAB, MaxQDA, user developed algorithms and open source visualization tools.

2.2.4 Data Publication

This part of the interview wanted to see how researchers publish their data, in case they do, and to explore the reasons behind not publishing data at all.

- Although some researchers have used different national and international data centres to deposit their data (this is mostly the case of Social Science researchers depositing data at the Economic and Social Data Service) the vast majority of researchers interviewed had never published data in a domain specific data archive.

- Researchers in Oxford are in many cases using departmental websites to publish their data but *"at the end of the projects, who owns that data, who takes care of it, do we just destroy it?"*

- Most of the participants on the interviews agreed that if data is produced with public money it should be made publicly available so that others can take advantage of it. They were mostly aware of how expensive data collection can be and how *"data reuse is the green way of doing things"*

- In most cases, researchers interviewed saw the usefulness of having summary data embedded in published articles to be deposited in a format that allows manipulation.

- There are several reasons for researchers to not deposit their data but the main one is they are that this would involve some extra work and that funding agencies do not require it. In addition to this, researchers can be very attached to their data and not very keen on sharing it openly:

  "Data feels quiet personal because of the effort that it takes to collect it, I am happy to share with serious people who are going to do good work, not willing to share with everyone "

- Publishing some types of data poses many challenges. An example is the ethical clearance needed for accessing some of the medical data produced. As one participant argued: *"Who takes responsibility to deal with the ethical clearance?"*

- Some researchers expressed a need to get advice on their IP rights when publishing their data and specifically when the data is published for a fee to those in for-profit organizations.

2.2.5 Support

This section of the interview was designed to learn about the support researchers get to manage their data and where do they turn for help when they encounter problems.

- Researchers felt that none or very little support is provided to help them manage their data. The support available mostly comes from technical officers in the departments or members of OUCS and involves advice on options for storage and sharing and in some cases database design. In many cases, the support received depends on the contacts within the institution:

  "… the support I get is because I happen to know people that are responsible of different services but there is nothing available, if I was new to Oxford I wouldn't have a clue"

- Few research groups have dedicated Data Managers or IT specialist who take responsibility on multiple aspects of data management like designing interview questionnaires, databases, data input tools and looking after the secure storage and access to the data.

- When looking for help to manage their data researchers tend to go to technical staff rather than librarians as they see their problems as mainly technical. Nonetheless, in some cases researchers using secondary data would welcome assistance to locate and access data resources.

- In certain research groups they saw extremely important to have whatever support for doing the data management embedded in their group so that their specific requirements from researchers in this group can be understood. Not only that, the person responsible for managing their data will require relevant experience in their research area to be able to make any sense of them.

2.2.6 Top Requirements for Services

At the end of the interview challenges and worries, in terms of managing their data, were discussed with interviewees and they were asked to suggest services that could help them do their work more effectively. Most of these issues had mostly came up previously during the interview but this helped to capture their top requirements.

Support

On the ground researchers as well as Data Managers, IT Officers and Admin Staff require hands on assistance and advice in many aspects of data management. Some of them are staring to produce new forms of digital material and are not fully aware of best practice for managing those.

> *Advice on practical issues related to managing data across their 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.*

Infrastructure

Researchers in the divisions of Mathematical, Physical and Life Sciences and Medical Sciences who are producing simulation data, huge in size, and imaging data from laboratory instruments, also of large volume, are at the moment struggling to store and share them securely amongst collaborators.

> *A secure and user-friendly solution that allows storage of large volume of data and sharing of these in a controlled fashion way allowing fine grain access control mechanisms.*

The sustainability and long-term preservation is a requirement that comes from across disciplines but especially from those were domain specific data repositories do not exist and mainly from those interviewed in positions of Head of Department or Head of Divisions. As one of the interviewees said:

> *"What happens with the infrastructure and data once the funding is over? Supporting data and infrastructure at the end of the project cost money"*

> *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.*

Funding

Many researchers felt that current data management could be improved greatly with the appropriate investment in activities to support better practices within departments.

> *Funding that could help address some of the departmental challenges to manage the research data that are being produced.*

3. RESEARCH DATA MANAGEMENT WORKSHOP

The Research Data Management Workshop was held on the 13th of June at the Said Business School. There were 46 attendees throughout the day from 24 departments, research centres and colleges as shown in the table below. There were also representatives from the Joint Information Systems Committee (JISC), the UK Research Data Service (UKRDS), the Research Information Network (RIN), the Digital Curation Centre (DCC), the European Bioinformatics Institute and the NERC Centre for Ecology and Hydrology. The workshop programme can be found in appendix 2.

Departments	
Admin	Man Institute of Quantitative Finance
African Studies Centre	National Perinatal Epidemiology Unit
Chemistry Research Laboratory	New College
Department of Clinical Pharmacology	Nuffield College
Department of Education	Oxford e-Research Centre
Department of International Development	Oxford Internet Institute
Department of Materials	Oxford University Computing Services
Department of Pharmacology	Oxford University Library Services
Department of Zoology	Physics Department
Educational Studies	Radiobiology Research Institute
Institute of Archaeology	St Edmund Hall
Jesus College	Wellcome Trust Centre for Human Genetics

Table 3. Attendees' departments

The top requirements for services to help researchers manage their data resulting from the interviews were presented and there was a general agreement on those. During the day it was made clear that there are many pockets of expertise on data management across the University but these are not co-ordinated and it is hard for researcher to know who to ask for advice. For this reason it was felt that researchers would found very useful a single point of contact for their data management queries. Feedback from the workshop suggested that participants liked particularly the discussion at the end, the range of examples shown and hearing about data management issues from researchers. Concerns expressed included: lack of representation from the Humanities in the programme; better representation of funding bodies; more information on IP issues, more practical examples tailored to the different divisions and available support.

3.1 Researchers' Presentations and Case Studies

The workshop included talks from Oxford researchers explaining their data management practices and challenges. Those talks form the basis for the following case studies.

CASE STUDY A – The National Perinatal Epidemiology Unit (NPEU)

BACKGROUND

The National Perinatal Epidemiology Unit (NPEU) is a research unit within the University of Oxford funded by the Department of Health in 1978. Its multidisciplinary research team conducts research dedicated to improving the care provided to women and their families during pregnancy, childbirth and the postpartum period, as well as the care provided to the newborn.

RESEARCH DATA COLLECTED

The NPEU collects and uses clinical trials data that is gathered by health professionals and from participants both collected in paper format through questionnaire forms. These data are highly sensitive as they contain individuals' personal information. Once collected, the data are then input into databases that are then securely stored, and indefinitely, at the Information Management Service Unit (IMSU). The data gets annotated with a description of the dataset with all the variables and information about the protocols and any ethical approval. Data policies are in place for data collection and access and these are circulated amongst new members of the unit.

Their data is highly valuable and it is important to preserve them. This is prove by the example of a drug administered to women in early pregnancy in the eighties to help them prevent miscarriages and how that drug caused vaginal cancers in the daughters and is now causing vaginal cancer in the daughters of the daughters. This demonstrates how clinical trials data lasts beyond the life-time of the individuals who are exposed to it and there may be the need to go back to the data to looks what has happened to subsequent generations.

CHALLENGES

They have been collecting data from 1978 and some years ago they had to migrate 113 datasets in between platforms. This took three months work from a senior programmer plus the continuous effort to ensure that current versions are still readable.

As a clinical trials unit, the NPEU has the objective to be registered as a high quality trials unit by the UK Clinical Research Collaboration (UKCRC) to gain future funding to run trials. Sound data management procedures and policies is a key competency they need to prove to form part of the UKCRC register.

When other researchers ask them access to their data many open questions arise. Who owns the data? Some of them were collected 30 years ago and the principal investigators are no longer there. What are the valid protocols for sharing these highly sensitive datasets? Can these data be transferred across countries?

REQUIREMENTS FOR SERVICES

Peter Brocklehurst, Director of the NPEU, indicated that the top two requirements for services to help the NPEU manage their data are:

- A central storage facility either nationally or institutionally that would allow them to securely store their data and not worry about security and preservation.
- Guidelines for standards for creating, documenting, accessing and sharing data.

CASE STUDY B – A researcher in the Department of Clinical Pharmacology

BACKGROUND

Simon Briggs a postdoctoral researcher in the Department of Clinical Pharmacology, his background is biotechnology and he has a PhD in Chemistry. He is a member of an applied translational science group within the Gene Delivery Group generating vast amounts of data.

RESEARCH DATA COLLECTED

His research process involves collecting data from laboratory instruments, see figures below, that contain some information like the set-up of the instrument, date and time but the files are in proprietary formats. These data then get stored on his computer and are manipulated using various statistical analysis and graphical manipulation tools. The processed data are then used in conjunction with the methodology captured in the lab notebooks so that he can publish a paper.

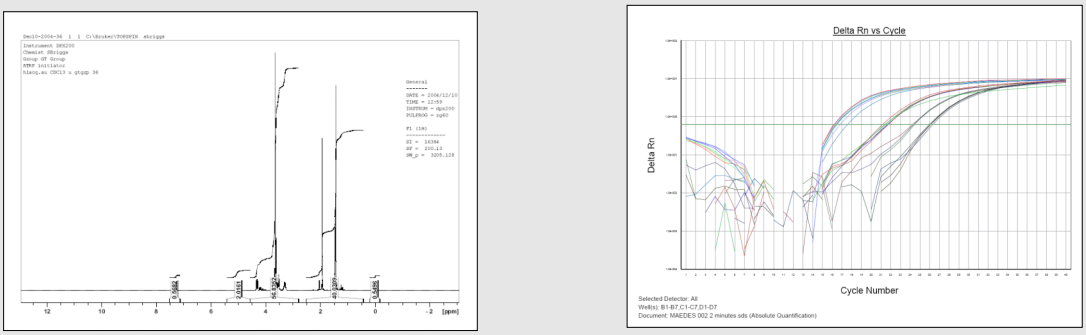

Figure 2. Research data generated by instruments

CHALLENGES

During his PhD he encountered many problems to file and organize his research data as he has not have any training on data management. Some years after completing his PhD he tried using the data he collected and it took him months to make sense of them.

Whilst working on the Department of Clinical Pharmacology they have trial a knowledge management solution to organize and share their data. He has realised that they are dealing with multiple data formats; disparate data types – chemical, biological & physical types; no access to metadata contained within files due to proprietary data formats; lack of integration of electronic records of the experimental procedures; people don't like to share the information on their computers; visualising and analysis of all the data for a single experiment is often best done on a big desk with lots of pieces of paper.

REQUIREMENTS FOR SERVICES

From Simon's point of view, any service for dealing with research data needs to be aware of the importance of preserving not only the data but also the methodology of the experiment that created the data in the first place. It is also important establishing the link between the data and existing and forthcoming publications that use the data. Finally, Simon believes that for any such activity to succeed, it needs to be implemented from above and there needs to be incentives for the principal investigators.

CASE STUDY C – Managing qualitative and codified data in Social Sciences

BACKGROUND

Joerg Friedrichs, from the Department of International Development in Oxford, engaged in a research project to answer the question: What makes states willing or unwilling to engage in international police cooperation? As a result of this project he published the book: Fighting Terrorism and Drugs: Europe and International Police Cooperation (Routledge, 2008).

RESEARCH DATA COLLECTED

To answer the previous research question, Joerg used codified data and simple descriptive statistics in an attempt to control complexity from 48 case studies. The data was collected from policy statements of key decision makers and then codified using a coding schema. Initially the data were stored as MS Word documents and later on as MS Excel files.

Once the book had been published, the data was also made openly available on the web using a web hosting service. The website represents a special service to readers of the book and provides codebooks and extra background information. Joerg got assistance from a skilled student who worked in close collaboration with him to convert the excel files into a dynamic website [http://joerg-friedrichs.qeh.ox.ac.uk/]

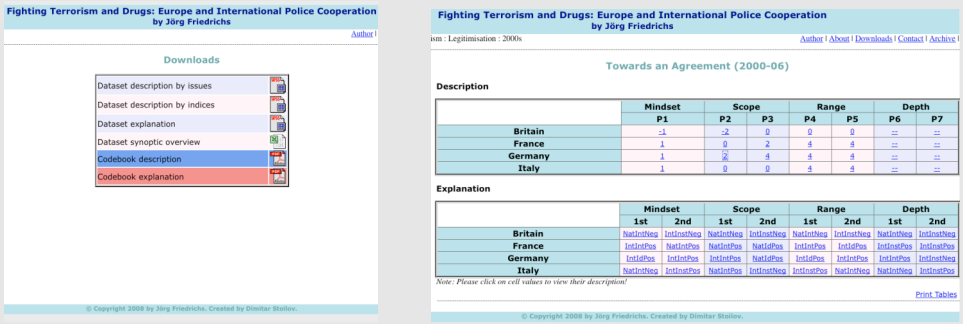

Figure 3. Screen-shoots from website with data complementing book. The cells of the table on the right are linked to a popup window containing the qualitative reasons for coding.

The website also contains archival sources which have also been deposited at the Historical Archives of the European Union [http://www.iue.it/ECArchives/EN/]

CHALLENGES

Joerg has to pay for the web hosting service [www.joerg-friedrichs.de] and the performance of this is not the desired one. He also provides access to these data through the departmental website. He believes that the optimal solution to his case would have been the editor (Routledge) publishing his data.

REQUIREMENTS FOR SERVICES

He considers that one-stop repository might have been helpful but was concerned that bureaucratic hurdles would have led to this qualitative data being pressed into a template. The two services that would see as most helpful to his type of research work are:

- Data management expertise offered on request for specific problems.
- Funding for autonomous projects.

3.2 National Initiatives

Some national initiatives relevant to Oxford efforts to manage and curate research data were also presented at the workshop and findings from those will be used in this project to form the final set of recommendations.

Simon Coles showed with examples how new generations of chemist exchange chemical information in new ways thanks to tools on the web such as blogs and pubcasts (video plus paper). He highlighted the importance to capture, store and manage this information and suggested institutional repositories and the electronic laboratory notebook as possible solutions with the recently created OAI-ORE[6] protocol to aggregate and exchange content.

ArrayExpress microarray is one the core services of the European Bioinformatics Institute providing a data service for array based transcriptomic data from the biomedical community. Helen Parkinson provided an overview of the service including some information about standards, their curation practices and data access models.

An update on progress twelve months after the "Dealing with Data" [7] report presented by Liz Lyon suggested that most of the recommendations had been taken forward through initiatives like the JISC Data Audit Framework projects[8], the e-Crystals Preservation and Curation Study[9], the UK Research Data Service (UKRDS) feasibility study[10]and the DCC Research Data Management Forum[11]. Nonetheless, at the national level the join up of funding bodies and policy makers to provide a national strategy for research data has not happened.

Stéphane Goldstein from the Research Information Network (RIN) presented their recently published report on the publication and quality assurance of research data outputs[12] where they investigate whether researchers make their data available to others, how they do this and if not why not. This report suggests that policy makers should take full account of the different types of data; there is a need for cooperation amongst funders, institutions and researchers; both funders and institutions should promote data dissemination and reuse; there is scope for publishers to promote ease of access and use of research data.

Robin Rice introduced DISC-UK DataShare[13] and the Data Audit Framework, two JISC projects related to the management and curation of research data in which the University of Edinburgh is involved. The former, which includes Oxford as a partner, aims at establishing new models for data archiving and sharing using repository technology whilst the latter is an implementation of a version of the Data Audit Framework to help institutions to map existing data collections, policies and practice in data curation.

[6] OAI-ORE (Open Archives Initiative-Object Reuse and Exchange) www.openarchives.org/ore/
[7] www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf
[8] http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2007/dataauditframeworkpilots.aspx
http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2007/dataauditframework.aspx
[9] www.ukoln.ac.uk/projects/ebank-uk/curation/eBank3-WP4-Report%20(Revised).pdf
[10] www.ukrds.ac.uk/
[11] http://www.dcc.ac.uk/events/data-forum-2008/
[12] http://www.rin.ac.uk/data-publication
[13] http://www.disc-uk.org/datashare.html

Finally, a four months JISC study on research data preservation cost[14] presented by Neil Beagrie showed how the costs for research data repositories are an order of magnitude greater than those for typical institutional repository focused on e-publications alone. The study also proposes a cost framework based on Full Economic Costings (FEC), tailored to research data which takes into consideration archive economics and *first-mover innovation* costs.

4. CONCLUSIONS AND NEXT PRIORITIES

Both the scoping study interviews and the workshop have been extremely useful to understand current data management practices and researchers' requirements for services. In addition to this, the previous activities have encouraged internal discussion on best ways forward for managing data in Oxford.

The priorities of the project for the next months include the following deliverables: a consultation exercise with support services available in Oxford, the organization of a second workshop and the production of a set of recommendations for digital repository services for research data.

As it has been mentioned previously, there are many services already available scattered across the University with expertise on data management but researchers are not finding particularly easy to discover those. The consultation exercise with service providers in Oxford, see table below, is aimed to validate the requirements identified and find out about their current services on offer that could assist researchers with their data management and plans for future ones. This will help to obtain a comprehensive picture of the current and planned infrastructure to support researchers with their data as well as to assist to identify gaps in the provision of services.

Support Services
Oxford University Computing Services
Oxford University Library Services
Research Services Office
Oxford e-Research Centre
Information Management Service Unit
Nuffield Social Science Data Library

Table 4. Support services available in Oxford

The second workshop of the project will be designed and organized in mid October to raise awareness amongst Oxford researchers of the support services already available as well as to inform about other relevant research data management related initiatives and developments in the UK.

The findings from the scoping study interviews, the consultation exercise with Oxford service providers and the two workshops as well as the results from the UKRDS feasibility study will assist in the production of the set of recommendations to improve and coordinate the provision of digital repository services for research data in Oxford.

[14] http://www.jisc.ac.uk/publications/publications/keepingresearchdatasafe.aspx

APPENDIX 1 - INTERVIEW FRAMEWORK

The following framework is based on the interview frameworks developed for the IBVRE and eIUS projects with some changes to adjust it to the aim and objectives of this scoping study.

Introduction

Give brief introduction to the Scoping Digital Repositories Services for Research Data Management including overall aim and objectives. Provide an overview of the questions that will follow and remember the interviewee about the nature of the semi-structured interview, the intention of taking notes, record the interview (with permission) and to publish findings.

Interview

1. Could you explain briefly what is your area of research and the types of research questions, with examples, that you try to answer?

2. I am interested in learning more about the research tasks that involve some form of data management that you carry out in order to help you move forward with your research agenda. I'm interested in doing this by going through one of your research projects in the context of a generic "research life-cycle", from funding application, data collection/processing, all the way to publishing to understand to what extent the following elements fit in your average working day.

a. The funding application – increasingly funding agencies require data management and data sharing plans as part of the funding application.
   - When applying for funding how do you decide that new data will need to be collected and how do you go about providing a plan for this?

With this question I want to learn more on how researchers think about data at this stage, why they decide that data needs to be collected, how they ensure that this data has not been created already and how they go about making data management plans.

b. Data collection –
   - Could you please explain what sorts of data (primary, secondary, experimental, simulation) do you collect and provide details about the process of collection?

In this part my aim is to engage in conversation to find out about data collection methods, types of data produced, the instruments and software used to do this and whether the data could be helpful to others. I will also ask whether secondary data is used to find out where and how are found and accessed. Finally I will explore why the collection of data happens in the way described (is it a discipline or departmental common practice?)

c. Processing of data –

- Once the data have been collected could you describe how they get processed i.e. how they get annotated, where are they stored, what security measures are taken to preserve confidentiality or integrity, etc?

Here I want to make sure that I understand how annotation/storage/back-up/manipulation/analysis/collaboration happens. Again, I will explore why the processing of data happens in the way described (is it a discipline or departmental common practice?)

d. Publishing – the publication of the research outputs is the end of this generic "research life-cycle", what happens with the data after this i.e. they get published or deposited somewhere, you need to destroy the data, etc?

In this part of the life cycle I want to find out whether deposit in archive occurs and if not I will attempt to find out the reasons that stop researchers doing so (data needs to be destroyed, does not want to share initially or at all, no place to deposit, etc) and where will the data be stored.

3. How are researchers supported either at local or institutional level for carrying out all the management of data required?

With this question I will attempt to figure out how support for data management across the generic life cycle occurs (researchers help each other at local level, departmental guidelines, etc).

4. What are your challenges and worries when managing research data and what services would help you do this work more effectively?

With this question I will attempt to get a top 3 requirements for services that would be most useful to researchers.

5. Is there anything else that you would like to add?

De-Brief

6. How do you think the interview went?

7. What are the benefits you believe you get from participating?

8. Could you suggest anyone you know that could participate in these interviews?

APPENDIX 2 – RESEARCH DATA MANAGEMENT WORKSHOP PROGRAMME

Abstract

The workshop is organised under the umbrella of the Scoping Digital Repository Services for Research Data Management project. The overall aim of the workshop is to hear examples of good and interesting practice, from Oxford and elsewhere, with respect to the use of digital repository services at various points in the research lifecycle, and from the perspective of various discipline areas. The event is designed for Oxford researchers to learn about best practice in research data management across disciplines and to encourage discussion.

Objectives

- Raise awareness of the benefits and importance of actively managing research data
- Communicate significant developments in data repositories in order to stimulate the adoption of best practice
- Complement the findings of the scoping study interviews with Oxford researchers

Programme*	
9.00	**Welcome and introduction** • Paul Jeffreys, Director of IT at the University of Oxford
9.15	**Scoping Digital Repository Services for Research Data Management** • Luis Martinez-Uribe (Oxford e-Research Centre, University of Oxford)
9.45	**Examples of research data management** • ***Archiving electronic data: An example from the NPEU*** Peter Brocklehurst (National Perinatal Epidemiology Unit, University of Oxford) • ***Data and knowledge management: A research scientist's perspective*** Simon Briggs (Department of Clinical Pharmacology, University of Oxford)
10.45	Refreshments
11:00	**National initiatives** • ***Dealing with Data - Perspectives on Progress to Date*** Liz Lyon (Digital Curation Centre) • ***The meaning of data "publication"*** Stephane Goldstein (Research Information Network) • ***DISC-UK DataShare and Data Audit Framework Implementation*** Robin Rice (Edinburgh University Data Library and EDINA) • ***Keeping Research Data Safe: A Cost Model and Guidance for UK Universities*** Neil Beagrie (Principal Consultant)
12.30	Lunch
14.00	**Examples of research data management II** • ***Supporting Capture, Preservation and Dissemination of Chemical Data*** Simon Coles and Jeremy Frey (University of Southampton) • ***'At the sharp end' - managing and extracting biological knowledge from high throughput biological data'*** Helen Parkinson (European Bioinformatics Institute) • ***Tailor-made: managing qualitative and codified data of an explorative research project on the sources of state preferences*** Joerg Friedrichs (Department of International Development, University of Oxford).
15.30	Refreshments
15.45	**Round Table Discussion -** (45 mins)
* Presentation slides and audio recordings available at: http://www.ict.ox.ac.uk/odit/projects/digitalrepository/Workshops.xml	