USING THE DATA AUDIT FRAMEWORK: AN OXFORD CASE STUDY

SCOPING DIGITAL REPOSITORY SERVICES FOR RESEARCH DATA MANAGEMENT

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

Author	Luis Martinez-Uribe (luis.martinez-uribe@oerc.ox.ac.uk) Digital Repositories Research Co-ordinator
Document name	DAF-Oxford.doc
Version	2.1
Date created	19/09/08
Date last modified	Last saved by Luis Martinez-Uribe on 24/3/09 11:21
Distributed to	

A collaborative project between

OFFICE OF THE DIRECTOR OF IT
Enabling Oxford University to make optimal use of IT

Oxford University Computing Services

oerc Oxford e-Research Centre

Oxford University Library Services

Oxford Digital Repositories Steering Group

funded by

DataShare project

JISC

JISC

Project Document Cover Sheet

DataShare project

Project Information			
Project Acronym	DataShare		
Project Title	DISC-UK DataShare		
Start Date	March 2007	**End Date**	March 2009
Lead Institution	EDINA, Edinburgh University		
Project Director	Peter Burnhill, Mark Brown		
Project Manager & contact details	Robin Rice, Edinburgh University Data Library, Main Library Bldg., George Square, Edinburgh EH8 9LJ R.Rice@ed.ac.uk, 0131 651 1431		
Partner Institutions	Universities of Edinburgh, Oxford, Southampton and London School of Economics		
Project Web URL	http://www.disc-uk.org/datashare/		
Programme Name (and number)	JISC Repositories and Preservation Programme: Repositories Start-up and Enhancement projects strand		
Programme Manager	Andrew McGregor		

Document Name			
Document Title	Using the Data Audit Framework: An Oxford Case Study		
Author(s) & project role	Luis Martinez, University of Oxford, Project Officer		
Date		**Filename**	
URL	http://www.disc-uk.org/publications.html		
Access	☐ Project and JISC internal	☐ General dissemination	

Document History		
Version	**Date**	**Comments**
1	24/03/09	

Acknowledgements

I would like to thank everyone that helped me collecting the information for this report and especially Anne Yates and Alan Garny for devoting their time and effort to provide me with very useful information about their research groups. Thanks also to Harry Gibbs from the University of Southampton for proofreading and suggesting changes to the report.

Contents

1. INTRODUCTION

The project Scoping Digital Repository Services for Research Data Management started in January 2008 as a cross-agency collaborative effort in Oxford. The project aimed to scope the requirements for digital repository services to manage and curate research data generated by Oxford researchers. The project contributed to the HEFCE funded UK Research Data Service feasibility study.

As part of the requirements gathering exercise around 40 interviews with researchers took place and a consultation with service units in Oxford was conducted. The interviews with researchers helped us to learn more about their data practices and to capture their top requirements for services to support their data management. The top requirements included:

I. A sustainable infrastructure that allows publication and long-term preservation of research data for those disciplines not currently served by domain specific services;
II. A secure and user-friendly solution that allows storage of large volumes of data and sharing of these with fine grained access control mechanisms;
III. Advice on practical issues related to managing research data across the research life cycle.

The consultation with service providers used the data management and curation services framework[1], shown in figure 1, to understand what services are available and identify gaps in the service provision.

Figure 1. Research Data Management and Curation Framework

The results of this consultation showed how expertise is widespread amongst service units in Oxford but on the whole, the vast majority of the research data management and curation services identified are not being offered fully or at all by service units across the University.

The application of the Data Audit Framework (DAF) Methodology in Oxford was undertaken as part of the JISC funded DISC-UK DataShare project, in order to identify data assets within a selection of research groups and pilot the methodology in Oxford.

This report is organised as follows: a brief description of the DAF methodology; explanation of how the methodology was adapted to be used in Oxford; description of the results from the two research groups and the study of the research data published on the Oxford website. The final section discussed some of the issues encountered when using the DAF methodology, the register and the online tool.

[1] More information about the Research Data Management and Curation Framework can be found at http://oxdrrc.blogspot.com/2008/12/research-data-management-services.html.

2. THE DATA AUDIT FRAMEWORK METHODOLOGY

The Data Audit Framework project was funded by JISC as a result of one of the recommendations made by the report Dealing with Data[2] to produce a methodology providing a framework to enable UK Higher Education Institutions to carry out audits of departmental data collections, awareness, policies and practice for data curation and preservation.

Led by the Humanities Advanced Technology and Information Institute (HATII) at the University of Glasgow in association with the Digital Curation Centre (DCC), a methodology, an online tool and a registry were produced. Four institutions piloted the methodology in order to test it and promote its uptake: University College London, Imperial College London, University of Edinburgh, and King's College London.

The Data Audit Framework Methodology consists of four main stages, see figure 2, that are designed to be run sequentially:

1. **Planning the audit**: Firstly an auditor is appointed, departments or research groups are identified and contacted to participate in the audit. After this, the audit is planned and an initial gathering of information takes place.
2. **Identifying and classifying data assets**: Here, the information collected initially is analyzed and interviews are arranged with relevant people to tease out the details about data management practices and data assets available.
3. **Assessing the management of data assets**: In this stage, assets are classified using the DAF audit forms to collect further information about each of the vital data assets identified in the previous stage.
4. **Reporting findings and recommending change:** The final stage involves producing a final report to provide the management with enough information to prepare a business case to justify an investment to deal with the data issues encountered.

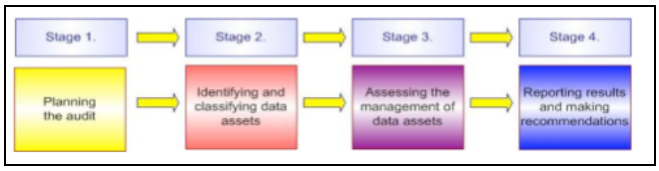

Figure 2. Four stages of audit framework

The auditor gathers information about the data assets that currently exist, where they are located, how they have been managed, which need to be maintained in the long-term and whether current practices put the data at risk. For this purpose, the auditor uses two forms. The audit form shown below, figure 3, contains basic information about the data assets such as name and description and will also help to classify them according to a given three-tier classification.

[2] Liz Lyon (2007) Dealing with data, available at http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2005/dealingwithdata.aspx

Audit Form 2: Inventory of data assets						
Name of the data asset	**Description of the asset**	**Asset Manager(s)**	**Reference**	**Classification**	**Classification comments**	**General comments**
Official name	*Basic description of content*	*Name or position*	*Where reference to asset was found*	*Choice of: Vital Important Minor*	*Reason for classification or comment on the classification chosen (should be based on the discussion with organisation's management) Could also include action suggested for each asset*	*Could include creation date, original purpose, type of data asset, or file format.*

Figure 3. Inventory of data assets, audit form 2

Once the first basic form has been completed, the form below is completed for those data assets of vital importance. This form provides more information about each data asset and how the management is exercised. There is a more detailed form that provides an extended set of optional information.

Audit Form 3A: Data asset management (Core element set)		
No	**Parameter**	**Comment**
1	**ID**	*A unique identification assigned by the audior or organisation to each data asset*
2	**Data creator(s)**	*Person, group or organisation responsible for the intellectual content of the data asset*
3	**Title**	*Official name of the data asset, with additional or alternative titles or acronyms if they exist*
4	**Description**	*A description of the information contained the data asset and its spatial, temporal or subject coverage*
5	**Subject**	*Information and keywords describing the subject matter of the data*
6	**Creation date**	*The date(s) on which the data was collected or created*
7	**Purpose**	*Reason why the asset was created, intended user communities or source of funding / original project title*
8	**Source**	*The source(s) of the information found in the data asset*
9	**Updating frequency**	*The frequency of updates to this dataset to indicate currency*
10	**Type**	*Description of the technical type of the data asset (e.g., database, photo collection, text corpus, etc.)*
11	**Format**	*Physical formats of data asset, including file format information*
12	**Rights and restrictions**	*Basic indication of the user's rights to view, copy, redistribute or republish all or part of the information held in the data asset. Access restrictions on the data itself or any metadata recording its existence should also be noted*
13	**Usage frequency**	*Estimated frequency of use and if known required speed of retrieval to determine IT infrastructure and storage needs*
14	**Relation**	*Description of relations the data asset has with other data assets and any any DOI ISSN or ISBN references for publications based on this data*
15	**Back-up and archiving policy**	*Number of copies of the data asset that are currently stored, frequency of back-up and archiving procedures*
16	**Management to date**	*History of maintenance and integrity of the data asset e.g. edit rights / security, and any curation or preservation activities performed*

Figure 4. Data asset management, audit form 3A

3. THE DAF METHODOLOGY APPLIED IN OXFORD

The JISC funded DISC-UK DataShare project funded the University Oxford to pilot the DAF methodology and share the experiences with the rest of the JISC community. The overall aim of applying the DAF methodology in Oxford was to map research data resources within Oxford at various levels on a group of selected research departments and institutes. Some of the objectives of the project included:

- To gain experience in using the DAF to map data resources and data management practices in departments and research groups across the University of Oxford.
- To generate an inventory of data assets and a basic appraisal for future decision-making about curation of research data.
- To identify valuable data assets that are being shared on public websites but could benefit from being saved in a plain format with enhanced metadata for enhanced discovery and long-term preservation.

3.1 Methodology

Stage 1 - Planning the audit

The Scoping Digital Repository Services for Research Data Management project had conducted a series of one to one semi-structured interviews with researchers across disciplines which aimed to learn more about their day-to-day work with data and capture their requirements for services to assist them with their data management duties. These interviews helped capture information about sources of funding, awareness of funders' requirements, data types created as well as data management activities throughout the research life cycle. Consequently, a part of the DAF methodology had already being carried out through the scoping study. The framework for these interviews can be found in Appendix 1.

The next step involved identifying the two research groups from different disciplines and data management cultures to participate in the Data Audit Framework pilot in Oxford: the Cardiac Mechano-Electric Feedback Group (CMEFG) from the Medical Sciences Division and the Young Lives Centre (YLC) from the Social Sciences Division. The selection of these was due to the existing good relationship after the scoping study interviews and the presence of staff with a Data Manager role in both groups. The Data Managers were contacted in order to see whether they would be willing to take part in the study. Furthermore, the scoping study interviews identified a sample of research data that was being made available on the Oxford University web pages and the DAF seemed to provide a good framework to collect valuable information about these datasets.

Stages 2 & 3 – Identifying and classifying data assets; assessing management of data assets

After the first contact, different meetings were arranged and emails were exchanged to explore and decide the best way to organize the gathering of information. The audit form in Appendix 2 was circulated to the main contact in CMEFG and this was returned with some of the information but several iterations were needed in order to have a complete and accurate picture of the data assets in that research group. The Data Manager of Young Lives shared a detailed Young Lives Data Management report required by ESRC summarizing their procedures and plans for future data collections. This document provided most of the information about data assets but further communication with the Data Manager was required. For the research data available on the Oxford University website, the DAF

online tool was used to gather the information and communication was needed with some of the data creators/owners to clarify parts of the information.

3.2 The Cardiac Mechano-Electric Feedback Group

Background Information

The Cardiac Mechano-Electric Feedback Group (CMEFG) is based at the Department of Physiology, Anatomy and Genetics and currently comprises thirteen staff and two students. The team conducts research into the mechanisms and implications of cardiac mechano-sensitivity, from the sub-cellular level to clinical models, using 'wet' experimental and 'dry' computational techniques.

Data Management and Curation Lifecycle

The CMEFG generates, as part of a Biotechnology and Biological Sciences Research Council (BBSRC) funded project, a substantial amount of data from a variety of imaging modalities such as magnetic resonance imaging (MRI) and histological sectioning. The data generated provides unprecedented detail of cardiac anatomy, in the case of histology giving much greater insight into the biological function, and opening up the opportunity for completely novel approaches to the computational modelling of the heart.

The BBSRC Data Sharing Policy[3] states that:

> *"Researchers are expected to ensure that **data are maintained for a period of 10 years after the completion of the research project** in suitable accessible formats using established standards where possible such that the data can be made available on request in line with BBSRC guidance on good scientific practice."*

The data are generated from microscopes that take high-resolution pictures of heart slices. One full heart has so far been completed and it consists of around 1800 sections which take up more than 1.5 TB. This is just one heart and as part of the project several hearts need to be processed in this way. It is expected that in its final year the project will create around 13 datasets of about 3,750 files, the size of which is envisaged to grow from 3.5 TB to 9 TB.

The experimental data is produced on the laboratory computers which are not connected to the network and then transferred to a storage solution, a network attached storage (NAS) system within the department with 2.7 TB storage available and internal back up mechanisms. The data is organised using folders and grouped by animal species, dataset number and type of data.

Initially annotation of data occurs through paper based laboratory notebooks where information about the experiment is recorded. From there, once the data is copied into the NAS system, some annotation is added through a wiki so that the whole process is well documented. No standards are used here. The data are stored on the NAS system using hierarchical folders with *readme* files that contain information about the data in that folder.

As part of a collaborative research effort, CMEFG shares the MRI and histological data with the Computational Biology Group (CBG) within the Oxford Computing Laboratory. Due to the sheer size of the data and the lack of appropriate infrastructure to share such datasets, the server needs to be physically transported and the data copied over to the Computing Laboratory servers. The MRI and histological data serves CBG to create and use 3D heart models on the National Grid Service (NGS) high performance computing services.

---

[3] BBSRC' data sharing policy: www.bbsrc.ac.uk/publications/policy/data_sharing_policy.pdf

The data produced as part of this project is also being shared with collaborators internationally. At the end of the research project, the data will be made publicly available as required by BBSRC but there are no national data facilities available at the moment to deal with the dissemination and preservation of these datasets.

Mapping to the DCC Curation Lifecycle Model

Figure 5 shows how the data management activities at the CMFEG Group map to the Digital Curation Centre (DCC) Curation Lifecycle Model[4].

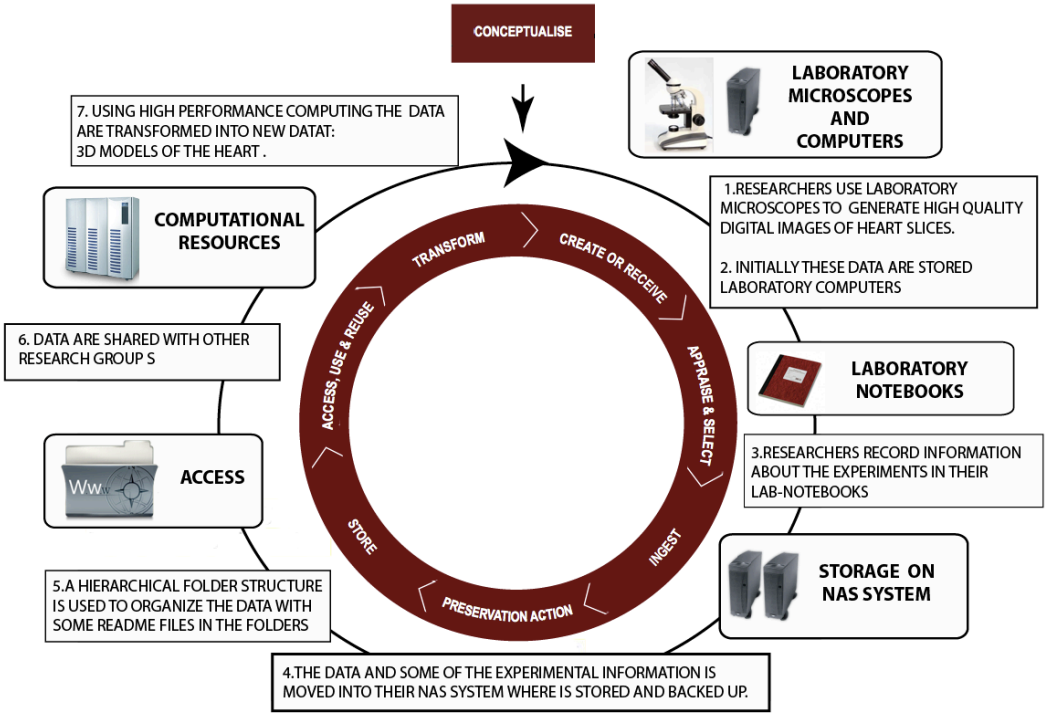

Figure 5. The CMEFG data management and curation workflow mapped to the DCC lifecycle model

3.2 The Young Lives

Background Information

The Young Lives research project is based in the Department of International Development in Oxford. Young Lives is an international study on childhood poverty in four countries over 15 years tracking the lives of 12,000 children in Ethiopia, India, Peru and Vietnam. Through interviews, group work and case studies with the children, their parents, teachers, community representatives and others, the project is collecting a wealth of information not only about their material and social circumstances, but also their perspectives on their lives and aspirations for the future, set against the environmental and social realities of their communities. This research may help answer research questions such as:

- How does poverty affect children's life-skills, education and their role in the family, and what are the opportunities and expectations they have for the future?
- What are the consequences of poverty for children's physical health and nutrition, cognitive abilities and educational progress, and their emotional and social well-being?

[4] The DCC Curation Lifecycle Model :www.dcc.ac.uk/docs/publications/DCCLifecycle.pdf

Data Management and Curation Lifecycle

The Young Lives project is funded by the Department for International Development (DFID). Although DFID does not have a data sharing policy, the project is required to disseminate the data produced as part of the Young Lives research for secondary analysis.

The project has a central Data Manager based in Oxford and one Data Manager in each of the four countries where the study takes place. This team is responsible for developing databases, monitoring data entry, running consistency checks, transforming the data to different formats for analysis in different statistical packages and disseminating the data.

Throughout the life of the project there will be five rounds of data collection, one every three years. In each data collection round, researchers conduct questionnaires and in-depth interviews with the children and their carers so they can build up a detailed picture of their daily lives. These activities create a wealth of quantitative and qualitative data. Two rounds of quantitative and qualitative data collection have already taken place and the third round of quantitative data collection begins in June 2009.

In rounds one and two of data collection, researchers in the UK and in each country designed the questionnaires and these were adapted according to the local needs. The data was collected using paper questionnaires in each location and inputted into MS Access databases. After the data is entered into MS Access databases a series of in country and oxford based consistency checks were done. Double data entry was also conducted to avoid any mis-keying of the data during entry. Once all the data had been collected and input, the databases were converted into SPSS using syntax files.

Data from rounds one and two were cleaned, i.e. anonymised from personal information, and it has been made publicly available[5] through the Economic and Social Data Service (ESDS) where high quality metadata, using the DDI[6] standard, is provided for the study. Registration with ESDS is required to access the data and researchers outside the UK need to apply for access.

After round one of data collection, a series of workshops and training events took place to standardise methods and expertise across the participating countries. These included different activities to coordinate the use of the different databases, double data entry systems and statistical packages. In addition to this, the main Data Manager prepared detailed quantitative data management guidelines to be followed in each country.

In round two, qualitative country teams generated large datasets in a variety of formats (field notes, audio recordings, photographs and videos). The Data Manager in Oxford prepared a document providing qualitative data management guidelines. These guidelines included information about the best formats, naming files according to given structures, guidelines for transcription, quality control, folder structure, backing up and explaining the role of the Data Manager.

During round one Data Managers in each country sent the data by email or posted CDs to the main Data Manager in Oxford. This method of transporting the data was unfeasible in round two due to the size of the databases. A web-based server was implemented to transfer the data across countries and only authorised users can upload and download the data through an encrypted password protected site. The data on this server are organised in folders

[5] The Young Lives: an International Study of Childhood Poverty: Rounds 1 and 2 are available through ESDS at http://www.data-archive.ac.uk/findingData/snDescription.asp?sn=5307&key=Young+Lives#doc

[6] The Data Documentation Initiative: http://www.ddialliance.org/

for quantitative and qualitative data and the four different countries. The methodology used for data collection and manipulation is also kept. Data back-up procedures are emphasised and the use of external hard drives is suggested throughout the data management guideline documents. The main server in Oxford is backed up using the University central service, the hierarchical file service.

Round three of data collection is scheduled to take place at the beginning of July 2009. PDAs will be used to collect the data for fieldwork and then they will be uploaded to databases adding a new element to their data management and curation lifecycle.

The Young Lives project website is also used to disseminate the project outputs and it provides information about the location of data and the publications using these data.

Mapping to the DCC Curation Lifecycle Model

Figure 6 below shows how the data management activities in the Young Lives project map to the Digital Curation Centre (DCC) Curation Lifecycle Model.

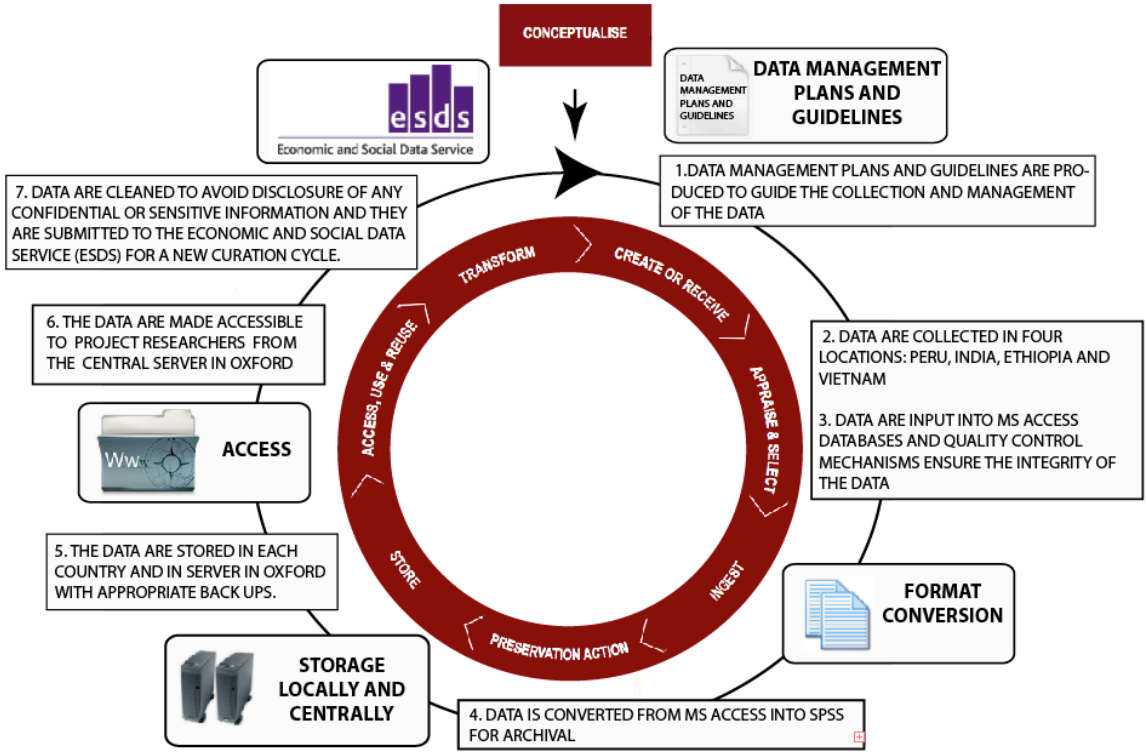

Figure 6. The Young Lives data management and curation workflow mapped to the DCC lifecycle model

3.3 Data Resources Available on the Web

During the scoping study interviews, researchers were asked whether they were publishing their data. Although some of the researchers interviewed were using national and international data centres to disseminate their data outputs, most of those who published their data were using departmental websites. The DAF register of data assets was applied to these data collections published on the Oxford website and discovered through the scoping study interviews and desk research conducted.

The DAF Online tool (http://www.data-audit.eu/tool/) was used to gather the information about the research data being made available on the Oxford University web pages. Some information about these datasets is provided in Appendix 3.

4. DISCUSSION OF RESULTS

The scoping study undertaken in Oxford before applying the DAF methodology helped to identify the research groups to work with and provided some initial information about their data management practices. This affected the form and use of the methodology which needed to be adjusted slightly to meet the objectives in Oxford.

Overall, the Data Audit Framework proved to be an extremely valuable methodology to plan and execute a strategy to gather information about data management activities and data assets held within research centres in Oxford. Some considerations about the methodology from the experience of applying it in Oxford include:

- More advice should be provided on how to identify research groups that generate valuable data. Maybe some examples could be provided of the different types of data that are expected to be produced by, for instance, researchers involved in clinical trials, crystallographers, social scientists doing field work, etc.
- Examples of interview frameworks and questionnaires to be used to gather evidence about data assets available and data management practices, like the one in Appendix 1, should be included in the methodology document.
- A template for reporting findings where the research data management practices in a given department could be ranked, subsequently providing a visual representation showing which data management areas need to be improved could be helpful when reporting findings to research departments.

The audit forms used to compile information about data and their management were very helpful. Some considerations about using the data audit forms include[7]:

- It has been difficult in some cases to decide what a data asset means. In the case studies it was very unclear whether a dataset was a file, images about one particular animal or the whole collection of images captured as part of the project.
- The classification of the data into vital, important and minor is not a very useful framework. More information is needed to help classify the data depending on its value and more information about how to assign value to data would be required for this.
- In the case studies it has been hard to decide who is the owner and who is the author of the data. It is very important to know the owners and authors to capture who funded the research and the principal investigators and tie the datasets to particular research projects in this way.
- The second case study in this report had data deposited in the Economic and Social Data Service (ESDS). ESDS data is well documented using DDI[8], a well-known standard for documenting social science data. It would be very useful to have mappings between accepted and widely used metadata standards for description and administration and the parameters suggested in audit form 3.

[7] Note here that the forms used in Oxford were obtained from a previous version of the methodology (1.3) and some of the following issues may have been resolved in the latest version.
[8] The Data Documentation Initiative (DDI) : http://www.ddialliance.org/

- In this report a mapping between the data management practices in the research groups and the DCC Curation Lifecycle Model has been provided but it would be of use to have a clear mapping between the data collected in the audit form 3 and the lifecycle model.

The online tool was only used for collecting data about the data resources available on the Oxford website. The experience with the tool has been positive but further work is required to add some extra functionality such as:

- Data exporting into a variety of different formats;
- Reporting tools informing on the parts of the audit that still need to be done, the number of datasets that have been discovered, etc;
- Searching tools;
- Common lists that everyone could use for some fields such as subject or data types;
- Capacity to add new fields as required by the auditor.

Appendix 1 - Interview Framework

The following framework is based on the interview frameworks developed for the IBVRE[9] and eIUS[10] 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 remind 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 briefly explain 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, in order 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 about 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) 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 about secondary data 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?)

[9] The integrative biology virtual research environment project :http://www.vre.ox.ac.uk/ibvre/
[10] The e-Infrastructure use cases and service usage models http://www.eius.ac.uk/

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 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 an 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 - Audit Forms

The Data Audit Framework (http://www.data-audit.eu/) is a methodology developed by HATII at the University of Glasgow and funded by the Joint Information Systems Committee. It is designed to enable research departments to carry out an audit of data collections and data management practices helping them to find out what data they hold, where is located and who is responsible for it. The form below has been designed to capture this information. Please fill in the form below for each data asset that:

- is still being created or added to;
- is used on frequent basis in the course of organization's work;
- underpins scientific replication e.g. revalidation;
- plays a pivotal role in ongoing research;
- or is being using to provide services to external clients and partners.

Audit Form	
Name of dataset	*Official name*
Description	*A description of the information contained the data asset and its spatial, temporal or subject coverage*
Owner	*Name or position*
Author	*Person, group or organization responsible for the intellectual content of the data asset*
Subject	*Information and keywords describing the subject matter of the data*
Date	*The date on which the data asset was created or published*
Purpose	*Reason why the asset was created, intended user communities or source of funding / original project title*
Source	*The source(s) of the information found in the data asset*
Updating Frequency	*The frequency of updates to this dataset to indicate currency*
Type	*Description of the technical type of the data asset (e.g., database, photo collection, text corpus, etc.)*
Formats	*Physical formats of data asset, including file format information*
Rights and Restrictions	*Basic indication of the user's rights to view, copy, redistribute or republish all or part of the information held in the data asset. Access restrictions on the data itself or any metadata recording its existence should also be noted*
Usage Frequency	*Estimated frequency of use and if known required speed of retrieval to determine IT infrastructure and storage needs*
Relation	*Description of relations the data asset has with other data assets and any any DOI ISSN or ISBN references for publications based on this data*
Back-up and Archival policy	*Number of copies of the data asset that are currently stored, frequency of back-up and archiving procedures*
Management to date	*History of maintenance and integrity of the data asset e.g. edit rights /security, and any curation or preservation activities performed*

Appendix 3 – Register of some data assets published on the Oxford website

Data Asset Name	Description
The SKADS Simulated Sky	A set of simulations of the radio sky performed at the University of Oxford, suitable for planning science with the Square Kilometer Array (SKA) radio telescope. http://s-cubed.physics.ox.ac.uk/
Fighting Terrorism and Drugs: Europe and International Police Cooperation	These data complement a published book. The analytical focus of the book is on the policy preferences of four European states: Britain, France, Germany, and Italy. The main research question of the book: What makes states in general, and large European states in particular, willing or unwilling to cooperate with other states in the fight against terrorism and drugs? The website provides tables with the descriptive and explanatory data as used in the book, including comments on the reasons for coding. All data were established in the context of a research project based at International University Bremen, Germany. http://joerg-friedrichs.qeh.ox.ac.uk/
LUPA: A new database on suicide missions in the Palestinian Area	The LUPA database provides data about suicide missions carried out in Israel, Palestine and Lebanon from 1980 to 2003. http://www.exlegi.ox.ac.uk/resources/Suicide_Missions/login.asp
Sicilian Mafia Dataset	The dataset, which is in Italian, was compiled between 1988 and 1992 using court files and confessions of Mafiosi who turned state witness, including the autobiographical book by an anonymous Mafioso. Using a grid of topics and concepts the sources were read and strings of text containing various items - such as descriptions of episodes or individuals, of modes and justifications of behaviours, interpretations of certain actions or relations - were picked from the sources and organised in 1103 indexed according to whether they fit one or other of the topics and concepts in the grid. These cards form the core of the dataset. By downloading the dataset and its supporting documents it is possible for the user to search and work with the data. http://www.exlegi.ox.ac.uk/data-service/browse-datasets.asp
FlyTED	FlyTED, the Drosophila Testis Gene Expression Database, is a public database currently containing 1,947 mRNA in situ images and ancillary data revealing the extent of expression of 623 individual genes involved in spermatogenesis in the testis of the fruitfly, Drosophila melanogaster, both in normal wild type flies and in seven meiotic arrest mutant strains. http://www.fly-ted.org/
The Beazley Archive	The original archive of Sir John Beazley, Lincoln Professor of Classical Archaeology and Art from 1925 until 1956, was purchased for the Faculty of Classics in 1965. There are now: - an estimated 500,000 notes - 250,000 black and white photographs - 33,000 negatives - 7,000 colour prints - 2000 books and catalogues -50,000 gem impressions http://www.beazley.ox.ac.uk/index.htm
Lexicon of Greek Personal Names	The Lexicon of Greek Personal Names was established in 1972 as a Major Research Project of the British Academy, at the suggestion of P. M. Fraser, Fellow of All Souls College, Oxford and a Fellow of the Academy. On acceptance of the proposal, Fraser was appointed Director of the project and Chairman of a supervisory committee. From the start, LGPN involved international collaboration, scholars from many countries being invited to contribute material and advice; but the Editors and central staff have always worked in Oxford. In October 1996, the project became part of Oxford University, under the aegis of the Faculty of Literae Humaniores, now the Faculty of Classics. It is a

	member of the group of Oxford Classics Research Projects. http://www.lgpn.ox.ac.uk/
Sphakia Survey	The Sphakia Survey is an interdisciplinary archaeological project whose main objective is to reconstruct the sequence of human activity in a remote and rugged part of Crete (Greece), from the time that people arrived in the area, by 3000 BC, until the end of Ottoman rule in AD 1900. Our research covers three major epochs, Prehistoric, Graeco-Roman, and Byzantine-Venetian-Turkish, and has involved the work of many people using environmental, archaeological, documentary, and local information. http://sphakia.classics.ox.ac.uk/index.html
The Thomas Gray Archive	The Thomas Gray Archive is a long-term research effort devoted to the study of the life and work of English poet Thomas Gray (1716-1771). It contains a host of primary and secondary resources, including electronic texts of Gray's complete poetry, a calendar of his correspondence, a digital library of images and audio-visual media, and finding aids to Gray MSS. The Archive is conceived as a structured platform for scholarly communication and collaboration and is developing as a living forum with the discussions, annotations, and contributions shared by the academic community. http://www.thomasgray.org/
Force Migration Online	Forced Migration Online (FMO) provides instant access to a wide variety of online resources dealing with the situation of forced migrants worldwide. Designed for use by practitioners, policy makers, researchers, students or anyone interested in the field, FMO aims to give comprehensive information in an impartial environment and to promote increased awareness of human displacement issues to an international community of users. http://repository.forcedmigration.org/
Malaria Atlas Project (MAP) Data	The Malaria Atlas Project (MAP) has been funded for five years by the Wellcome Trust, UK. MAP is a joint project between the Malaria Public Health & Epidemiology Group, Centre for Geographic Medicine, Kenya and the Spatial Ecology & Epidemiology Group, University of Oxford, UK with collaborating nodes in America and Asia Pacific region. The main objective of this project is to develop a detailed model of the spatial limits of Plasmodium falciparum and P. vivax malaria at a global scale and its endemicity within this range. The first stages of MAP were data acquisition and archive. The various mechanisms by which we have assembled the largest ever database of malaria parasite rate (PR) data are found on the data page. MAP intends to release in the public-domain all data collected during the project for which permission to disseminate has been granted. The first full release of the parasite rate data is scheduled for June 2009 to enable global searches to be comprehensive and time for our endemicity maps to be tested and reviewed. http://www.map.ox.ac.uk/MAP_data.html
Computational Biology Research Group – Collaborative Data Area	The collaborative data area of the Computational Biology Research Group provides access to data resources of ongoing and published projects that the CBRG has produced in collaboration with members of the Medical Sciences Division. http://www.molbiol.ox.ac.uk/data.shtml
HIRDLS Data	The High Resolution Dynamics Limb Sounder (HIRDLS) instrument provides measurements of temperature, trace constituents and aerosols from the middle troposphere to the mesosphere, with a key attribute of high vertical resolution. HIRDLS will also provide key measurements of atmospheric aerosols and cirrus clouds, as well as unique measurements of sub-visible cirrus. This website provides access to satellite atmospheric chemistry data generated with the HIRDLS. http://www.atm.ox.ac.uk/hirdls/data/index.shtml
Childhood Cancer Research Group	The Childhood Cancer Research Group provides access to: - The National Registry of Childhood Tumours (NRCT) - the largest population-based childhood cancer registry in the world - Great Britain Registrations 1975-2000 - The number of registered cases of childhood cancer is available for each year between 1975 and 2000. - Five Year Survival Rates - Percentage of cases in the NRCT that survived for five years or more from diagnosis

	http://www.ccrg.ox.ac.uk/datasets/requestdata.htm
The National Perinatal Epidemiology Unit (NPEU)	The National Perinatal Epidemiology Unit (NPEU) is a multidisciplinary research team 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. Their registers, surveys and cohorts are explained with the access methods on their website: http://www.npeu.ox.ac.uk/obsepi
Department of Social Policy and Social Work Demographic data	The Department of Social Policy and Social Work provide access to three datasets taken from a variety of sources: - 'Expectation of life at birth, males and females, developed countries, 1945-2007'. - 'Births outside marriage per 1000 live births, developed countries, 1945-2007'. - 'Total Period Fertility, developed countries, 1945-2007'. http://www.spsw.ox.ac.uk/research/groups/oxpop/demographics.html
ICRISAT Village Level Studies Data	This website contains a new panel data set, linking the data initially collected between 1976 and 1985 by the International Crop Research Institute for the Semi-Arid Tropics (ICRISAT) as part of the Village Level Studies to a further rounds in 2001 to 2007. http://www.economics.ox.ac.uk/members/stefan.dercon/icrisat/ICRISAT/usingthedata.html
Malaria GEN data	Access to individual-level genotype data is available by application to the MalariaGEN Independent Data Access Committee. The data are stored in the European Genotyping Achive. Access to data will be granted to qualified investigators for appropriate use. http://www.malariagen.net/home/science/dataaccess.php
CSAE datasets	The Centre for the Study of African Economies publishes some datasets on their site: * Ghana Firms * Ghana Macro Data * Ethiopia Firms * Ethiopia Households * Comparative Firm-level Data * Tanzania Firms * Ghana and Tanzania Urban Household Panel Surveys http://www.csae.ox.ac.uk/datasets/main.html
Oxford Centre for Population Research	The Oxford Centre for Population Research makes data taken from a variety of sources available through this website: http://www.spsw.ox.ac.uk/fileadmin/static/Oxpop/data.htm