JISC

Project Document Cover Sheet

Project Information			
Project Acronym	EIDCSR		
Project Title	Embedding Institutional Data Curation Services in Research		
Start Date	01 April 2009	**End Date**	31 December 2010
Lead Institution	University of Oxford, Oxford University Computing Services (OUCS)		
Project Director	Dr Michael Fraser		
Project Manager & contact details	Dr James A J Wilson Oxford University Computing Services 13 Banbury Road Oxford OX2 6NN james.wilson@oucs.ox.ac.uk 01865 613489		
Partner Institutions	n/a		
Project Web URL	http://eidcsr.oucs.ox.ac.uk/		
Programme Name (and number)	JISC Information Environment 2009-11 (12/08)		
Programme Manager	Neil Grindley		

Document Name			
Document Title	Final Report		
Reporting Period	01 April 2009 – 31 December 2010		
Author(s) & project role	James A J Wilson, Project Manager		
Date	February 2011	**Filename**	EIDCSR Final Report v.1.0.docx
URL			
Access	☑ Project and JISC internal	☐ General dissemination	

Document History		
Version	**Date**	**Comments**
V 1.0	28-Jan-11	First draft of report
V 1.1	14-Feb-11	Minor corrections by Steering Group members

Table of Contents

Acknowledgements

The EIDCSR project was funded by the JISC Information Environment Programme and supported by the University of Oxford. The core project team consisted of, at various times, Luis Martinez-Uribe, Dr. James A J Wilson, Asif Akram, and Tahir Mansoori, under the direction of Dr. Michael Fraser. Governance and additional direction was provided by the Project's Executive Board and the Project Working Group, comprising Professor Jeff Haywood, Neil Grindley, Dr. Michael Fraser, Professor Paul Jeffreys, Professor David Gavaghan, Professor Peter Kohl, Dr. Alan Garny, Sally Rumsey, Kathryn Dally, Michael Jubb, Chris Rusbridge, Ben O'Steen, and Dave Rowell. I would particularly like to thank Kathryn Dally and Sian Dodd at Research Services for their work on the Research Data Management Portal and University Policy, the researchers involved in the 3D Heart project for providing feedback when called upon, and the original Project Manager, Luis Martinez-Uribe, for his continued contributions to the project even after taking up his new position in Madrid. His help and advice have greatly aided continuity and made the job of picking up the project management responsibilities half-way through the project a lot easier than it would otherwise have been.

Executive Summary

The Embedding Institutional Data Curation Services in Research (EIDCSR) Project formed part of a programme of activities at the University of Oxford to develop an institutional infrastructure for research data curation. EIDCSR worked with a team of scientists generating and using very high resolution two-dimensional and three-dimensional images of hearts in order to understand and help improve their data preservation and curation processes. The Project was an intra-institutional collaboration involving the Computing Services (who led the work), the Research Services Office, the Bodleian Libraries, and the researchers themselves.

EIDCSR aimed to develop and join up elements of research data infrastructure at various points of the data lifecycle by improving research data workflows underpinned by the development of a University Data Management Policy. A particular concern of the project was to ensure that the research data being generated could be securely preserved and documented in such a way as to enable future access and re-use. This was to be achieved by ensuring that the research data itself could be archived to and retrieved from the secure Hierarchical File Server (HFS) hosted by the Computing Services, whilst the descriptive metadata could be stored and searched in the Digital Asset Management System (DAMS) being developed by the Bodleian Libraries. Although EIDCSR worked with specific research groups, the intention of the project was that the resulting infrastructure could ultimately be extended to other research disciplines.

The EIDCSR Project achieved its objectives in that there is now sufficient core infrastructure in place to enable the research groups involved in the 3D Heart Project to document their data, store it securely off-site, browse the documentation, and retrieve data from the long-term file store. In some respects the project achieved more than it had initially set out to, additionally developing data visualisation software and a researcher-facing data management web portal.

Notable outputs include:

- Research Data Audit and Requirements Analysis Report
- Draft University Research Data Management Policy
- Research Data Management Web Portal (http://www.admin.ox.ac.uk/rdm/)
- Digital curation workflow module, with metadata and archiving client
- Core research data metadata schema
- 'Workbench' large image visualisation software
- Cost model for future Secure Storage and Metadata Management Service

Work remains to be done to bring the archiving and search client software developed by EIDCSR up to a production service level. The costs of doing so, as well as the costs of running an envisaged Secure Storage and Metadata Management Service are estimated in this report. Although the draft University Data Management Policy has been completed and submitted to the University's Research

Information Management Sub-committee (RIMSC), we have not yet heard whether the Policy has been approved at the time of writing this report.

Background

The EIDCSR Project is part of a programme of research data management infrastructure development at the University of Oxford, inspired both by the institution's desire to look after its digital research outputs and the growing recognition that the real value of research data is not at present being fully exploited. In particular, issues surrounding digital data preservation and re-use have become increasingly pressing in recent years, as it has become evident that expensively-assembled data outputs, often vital to understanding (and verifying) published research papers, are not being securely kept, nor made easily available to other researchers who might wish to reinterpret the data or even apply it to new research questions unimagined by its original creators.

Awareness of the importance of data curation has been growing throughout the decade, with principles and guidelines for enabling continued access to research data being published by various governments and scientific organizations.[1] In 2007 UKOLN produced a report for the JISC which recommended amongst other things that 'Each higher education institution should implement an institutional Data Management, Preservation and Sharing Policy, which recommends data deposit in an appropriate open access data repository and/or data centre where these exist'[2] Most of the UK HE funding councils had already implemented data policies requiring researchers to submit data management plans with funding bids by 2008,[3] although anecdotally enforcement of these plans was and remains patchy. By 2009 the need for researchers themselves to be directly involved in managing their data was becoming widely recognized, as is evident from an editorial in Nature, calling for researchers to be trained in information management ('a discipline that encompasses the entire life cycle of data'): 'data management should be woven into every course in science, as one of the foundations of knowledge'.[4]

The University of Oxford's programme of projects to develop data management infrastructure has its roots in a cross-University committee formed in 2006 to coordinate the development of digital repositories within the University. The Oxford Digital Repositories Steering Group (ODRSG) was chaired by the Pro-Vice Chancellor for Academic Services, University Collections, and Research. The Oxford Research Archive (the University's repository for eprints and theses) reported into the Group, as did activities related to e-learning. The ODRSG identified priorities for digital repository development including: a) to ensure interoperability between existing and planned repositories (the Group preferred to speak of a 'federated institutional repository' rather than simply an 'institutional repository' in order to reflect Oxford's devolved and federated nature as an institution); and b) to better support the management and curation of research data. The latter priority was driven partly by the Research Councils and other funding bodies increasingly requiring data management plans as a condition of funding, partly by the recognition that few University services existed to support the management of research data, and partly by new opportunities for large-scale research enabled (or potentially enabled) through the e-science agenda. All three factors were of course inextricably linked.

The ODRSG motivated the funding of an internal project, 'Scoping Digital Repository Services for Research Data Management',[5] which sought to establish exactly what was required by researchers at the University and what roles the various service groups could and should take to meet those

[1] OECD, 'Principles and guidelines for access to research data from public funding', 2007, http://www.oecd.org/dataoecd/9/61/38500813.pdf; International Council for Science (ICSU) 'ICSU report of the CSPR assessment panel on scientific data and information', 2004, http://www.icsu.org/Gestion/img/ICSU_DOC_DOWNLOAD/551_DD_FILE_PAA_Data_and_Information.pdf; Her Majesty's Stationary Office (HMSO) 'Science and Innovation investment framework 2004-2014', 2004, http://news.bbc.co.uk/nol/shared/bsp/hi/pdfs/science_innovation_120704.pdf.
[2] Lyon, L., 'Dealing with data: roles, rights, responsibilities and relationships', 2007, p. 6. http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2005/dealingwithdata.aspx.
[3] The Digital Curation Centre (DCC) provide a summary of funders' data policies at http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies.
[4] 'Data's Shameful Neglect', Nature 461, 145. http://www.nature.com/nature/journal/v461/n7261/full/461145a.html.
[5] Scoping Digital Repository Services for Research Data Management Project. http://www.ict.ox.ac.uk/odit/projects/digitalrepository/index.xml.

requirements. A data management service framework was derived from the requirements and used to evaluate the current services provided at Oxford and identify gaps in provision. The requirements analysis highlighted researchers' need to have access to expertise, tools, and infrastructure to manage their data in compliance with funding bodies' requirements. Key requirements to enable more effective support for research data management 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 for the storage of large volumes of data, together with fine grained access-control mechanisms;
iii. Advice on practical issues related to the management of research data across the research life cycle, including the development and implementation of data management plans.

The scoping project also identified particular 'data rich' research groups whose research necessitated the creation, analysis, and sharing of large volumes of data, with funding agency requirements mandating the preservation of data beyond the completions of the projects they were working on.

The national and institutional strands of planning the came together in early 2009 thanks to the JISC's Information Environment programme, which provided the funding that enabled the work undertaken via the EIDCSR Project.

The research groups with whom the project has been working were already engaged upon a collaborative BBSRC-funded project of their own when EIDCSR began, entitled 'Technologies for 3D histologically-detailed reconstruction of individual whole hears' (henceforth referred to as the '3D Heart Project). The groups consist of: the Cardiac Mechano-Electric Feedback Group based within the Department of Physiology, Anatomy, and Genetics (DPAG), a team from the Department of Cardiovascular Medicine (DCM), and the Computational Biology Group (CBG) based within the Oxford Computing Laboratory (ComLab). The 3D Heart Project involves the creation of very-high resolution images of hearts, segmenting these images, meshing them, and using the resultant 3D models for in-silico experiments to assess responses to electrical stimuli. The BBSRC Data Sharing Policy 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". The 3D Heart Project was regarded as a good exemplar due to its mixture of its data preservation requirements (and lack of obvious means by which these requirements could be met), the large size of the data being generated (ultimately anticipated to be in the region of 9TB), and the fact that many issues arising from the project were felt to be generic across domains. The use of state-of-the-art technologies to generate very large quantities of data (and particularly imaging data) is widespread, and has the immediate consequence of how such data can be made accessible to the teams using it in their research, and to the general research community, as well as the long-term availability required by the research funder.

It was anticipated from the outset of the project that the infrastructure developed for the researchers on the 3D Heart Project should and would be extendible to meet more general research data curation needs within Oxford, as well as providing a practical case-study for other Universities considering developing their own infrastructure.

Since the commencement of the EIDCSR project in April 2009, a second JISC-funded data management infrastructure project has begun at Oxford, the Supporting Data Management Infrastructure for the Humanities (Sudamih) Project.[6] Sudamih builds upon the EIDCSR work, but focuses on different disciplinary concerns and is looking to develop different aspects of infrastructure: researcher training; and the development of an intuitive 'Database as a Service' system (DaaS). The DaaS will constitute a centrally-supported online tool for researchers to create and share databases. Many members of the staff assigned to EIDCSR are also working on Sudamih, to ensure that development is coordinated and tools and resources are shared where appropriate.

Besides the University and the JISC, the EIDCSR project was also associated with the UK Research Data Service (UKRDS), for which Oxford is a 'path-finder' institution.[7] The UKRDS has the objective

[6] Supporting Data Management Infrastructure for the Humanities (Sudamih): http://sudamih.oucs.ox.ac.uk/.
[7] UK Research Data Services (UKRDS): http://www.ukrds.ac.uk/.

of assessing the feasibility and costs of developing and maintaining a national shared digital research data service for UK Higher Education sector, so the progress of EIDCSR was of obvious interest.

Aims and Objectives

The EIDCSR Project sought to develop infrastructure to address selected elements of the digital curation lifecycle, by embedding policy, workflow, and sustainability solutions within three 'data rich' research groups. The aim was to join up existing institutional and departmental services, using an approach that could scale to address the more general data preservation challenges of research groups that generate, share, and (potentially) reuse research data. This was agreed to require an institutional collaboration bringing together research groups, the Bodleian Libraries, the Computing Services, and Research Services, each of whom have a role to play in supporting the passage of research data through the lifecycle. The project also sought to integrate existing research workflows with the University's long-term file store and the Fedora Digital Asset Management (DAM) system. All of this was to be underpinned by policy development. The project was therefore aligned both with Oxford's commitment to the digital curation of research data and consistent with the data management plan and registry of research data proposed for the UK Research Data Service (UKRDS) Pathfinder service.

EIDCSR intended firstly to better scope the curation and preservation requirements of the selected research groups, and then to address these requirements by embedding selected data management tools and infrastructure into the researchers' workflows, thus improving data curation practices at various stages of the data life-cycle. The project hoped thereby to produce an exemplar case study that would be extensible to other research groups both within and beyond the University of Oxford.

Project objectives included:

- Embedding institutional services for data preservation for two [later adapted to three] specific research communities that can then be expanded to other research disciplines;
- Introducing sheer curatorial[8] practices within research workflows in an attempt to add value to everyday scholarly work;
- Developing and coordinating institutional and service level policies and economic models for the management, preservation, and sharing of research data;
- Investigating the roles and responsibilities of service providers in Oxford to support their researchers with the management and curation of research data by developing a deep understanding of research workflows with data and how they may interface with institutional services.

Besides the broad objectives, EIDCSR also planned a number of more specific deliverables:

- A report detailing the requirements from the three research groups as well their data holdings and management practices, based on the DAF methodology;
- Development of digital curation workflow module integrating the research data lifecycle with the HFS Archive and Fedora DAMS. The software would be available under an open source software license and be accompanied by a technical report documenting the requirements and development process;
- A report on the strategic provision of filestore and access management infrastructure within a devolved environment for large quantities of research data (with the support of IBM);
- Development of selected institutional and service level policies for the management, curation and preservation of research data outputs;
- A report on the application of costing and sustainability frameworks developed by the JISC-funded Keeping Research Data Safe and LIFE projects as well as any further JISC-funded cost analysis projects; and sample cost models commensurate with the policy framework;
- A project website with information about the project, including a blog to describe day to day experiences, and RSS-supported bookmarks of relevant activities and publications;

[8] Term coined by Alistair Miles while working on the DCC SCARP project referring to embedding curatorial activities into workflow of those managing and creating the data, see: http://en.wikipedia.org/wiki/Sheer_curation.

- Two workshop reports together with publications and presentations in relevant journals and conferences;
- A final report, describing the process of implementing the exemplar preservation solution and making recommendations on how JISC should consider continuing work in this area.

The functionality that was anticipated by the end of the project included:

- Digital curation workflow module for metadata capture, secure, long-term storage and retrieval of data and accompanying metadata;
- Integration with discovery service to enable discovery of curated data based on metadata;
- Linking of metadata between publications within the Oxford University Research Archive and data (in both directions, based on outputs from the BID Project)

There were no significant changes to the initial aims and objectives of the project, although some of the deliverables needed to be reassessed over the course of the work. The initial intention to produce a report on the strategic provision of filestore and access management infrastructure had to be abandoned due to difficulties engaging experts in the relevant sections of IBM, our partner for this aspect of the work. Whilst this was a little disappointing, it did not impact upon the rest of the project. The feedback we received from the researcher requirements-gathering phase suggested that the project should help with the development of a visualisation and annotation tool, which could be integrated into existing workflows and support data management. This was not envisaged as one of the initial deliverables, but become one later. Furthermore, the budget assigned to policy development turned out to be sufficient to partly cover the development of an additional deliverable: the development of a University Data Management Web Portal, hosted by Research Services. This was recommended as a 'quick win' in the recommendations arising from the policy work.

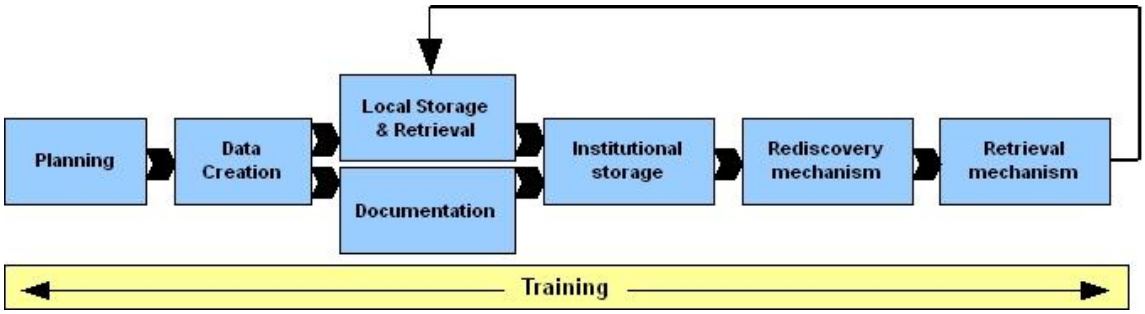
The extent to which the various objectives and deliverables of the project have been met by the project is addressed in the Outcomes section of this report.

Methodology

Approach to Institutional Data Management Infrastructure Development at Oxford

The University of Oxford has a highly federated structure, with the various academic divisions, the faculties within those divisions, and the supporting service groups each having a large degree of autonomy. This poses a challenge for developing a research data management infrastructure, as there is no one department in a position to 'own' the resulting set of services, nor enforce their use. Indeed, if the University is to achieve its objective of creating a solution for managing research data through all stages of the data life-cycle, then all of the different groups involved need to share a common understanding of the purpose of the enterprise and the processes required to make it work.

The data management challenge can be conceived as a sequence of steps each of which needs to be adequately completed in order that the data itself can progress to the next step, with the intention that it can ultimately be re-used, thereby maximising its value. If the infrastructure does not exist for any given step, or the agents involved do not understand what is required of them, then the potential value of the data cannot be fully realised.



The University of Oxford is taking the approach that when developing institutional data curation infrastructure it is vital to engage with researchers from the earliest stages of the data life-cycle: the research project planning stage (or 'conceptualisation' according to the DCC model).[9] It is, after all, the researchers who best understand the data they create, the processes by which it was derived, its purpose, and its limitations. Therefore it is the researchers who need bear ultimate responsibility for their data up until the point that it is archived and transferred to institutional storage.

There are already support services at the University of Oxford with responsibilities to help researchers at several stages of the sequence above, although at the start of the EIDCSR project there was little expertise or guidance available relating directly to data management and curation. EIDCSR aimed to strengthen several aspects of the chain, by engaging both researchers and the various support services, specifically Research Services, the Computing Services, and the Bodleian Libraries. Inevitably, it is the researchers themselves who have the most important role in the planning of new research, the creation and structuring of data, and the way in which it is organised and stored during the actual research process. Research Services, divisional and departmental administrators and research facilitators can however play an important role however in supporting the planning stage of research. The Libraries, being the traditional custodians of information, naturally have a role to play in providing data storage, rediscovery, and retrieval systems. The Computing Services at the University possess a very secure long-term electronic file-storage system, and have expertise in developing tools and providing services to support researchers manage and document their electronic data.

The general approach being taken to the development work at Oxford may be summarised as an attempt to advance towards a coherent infrastructure on all significant fronts, using the researchers to guide and validate each strand of development as it progresses. The research communities with which the projects are working would not claim any especial expertise in data curation. Indeed an important aspect of the projects is to be able to gauge the buy-in from 'normal' researchers whose focus is on the day-to-day research itself.

By working on several different aspects of data management in close coordination, the idea is that the project does not lose sight of the interrelated nature of data management activities. Unless preservation strategies can be allied with resource discovery services, for instance, the value of preservation is severely limited. Working simultaneously on various aspects of the data management chain also ensures that all of the relevant support services are involved and kept aware of the progress being made on other aspects – important to ensure buy-in and understanding of the 'big picture'.

Methodology applied by the EIDCSR Project

The EIDCSR Project broadly followed the three-stage process specified by the JISC in the Call for Proposals. This consisted of an initial analysis stage (gathering requirements), followed by a pre-implementation stage (translating requirements into policy and implementation plans), and then an implementation stage (during which existing and new services are further developed and integrated). The practicalities of ensuring each member of the project team was productively employed throughout the project life-span resulted in a slightly more agile *modus operandi* than this rather formal structure might suggest, with the technical development work in particular occurring in bursts of implementation interspersed with periods where it was necessary to go back to check requirements and modify implementation plans.

The Computing Services took the coordinating role during the project, holding regular meetings with representatives from the Bodleian Libraries, Research Services, and the researchers themselves. Each of the Services Groups was assigned particular responsibilities by the project to ensure that the most relevant expertise could be brought to bear upon each aspect of infrastructure development. Thus the Libraries were expected to advise on matters relating to metadata and to help integrate the metadata packaging and management services of their Fedora Digital Asset Management (DAM) system into the workflow. Research Services on the other hand was given responsibility for drafting an exemplar institutional data management policy, in accordance with both funding agency and institutional requirements.

[9] See http://www.dcc.ac.uk/resources/curation-lifecycle-model.

Implementation

The EIDCSR Project commenced in April 2009 and ran until the end of December 2010, after receiving a three-month no-cost extension.

The initial stage of the EIDCSR Project involved undertaking an audit of data assets and data management practices as well as a requirements analysis of the research groups involved in the 3D Heart Project. This work was based upon the Data Audit Framework (DAF) methodology and built on previous DAF work conducted in Oxford.[10] EIDCSR identified appropriate researchers to interview via a 'friend of a friend' approach, beginning with the co-PIs. Identified interviewees were then individually emailed, provided with the project brief, and invited to take part in the study. If they agreed, a short survey and the Data Asset Register form were circulated to them, which were followed up by interviews lasting for no longer than an hour. Eleven initial interviews were conducted. The Project Analyst used the interviews to develop a better understanding of the data produced and how it was used, and the approach taken to manage the data and what could be done to help manage the data more effectively in future. After each interview, the analyst produced a short summary and completed the Data Register Form as required by the DAF. The recorded transcriptions of the interviews were transcribed by a professional audio transcription company. The completed report was circulated to the Project Working Group and the Executive Board as well as the participating interviewees for approval and feedback. Some of the initial interviews were followed up later in the project to verify details, which was particularly important due to the change of project staff at the end of 2009.

The results and experience from this initial analysis stage fed into most of the other work packages such as the metadata management work package (wp3), technical implementation (wp4&5), policy development (wp7), and economics and sustainability (wp8). The outputs from these work packages are described in more detail in the next section.

The work package on data storage and access infrastructure (wp6) was independent of the others and assigned to our project partners at IBM. The output was planned to be a report on the strategic provision of filestore and access management infrastructure, although it later became apparent that the IBM researchers with whom we needed to engage were not those with whom we were initially in contact, and furthermore that they did not have the time to commit the resources we had anticipated to the project. Consequently, this work package was dropped from the project plan, although its abandonment did not have any impact on any of the other work being undertaken by the project.

A greater hindrance to project arose from the time taken to recruit a technical developer, which did not happen until January 2010. The technical work packages were therefore delayed by five months from their initial proposed starting date, and the implementation timetable adopted a slightly lob-sided appearance as the project attempted to make up for time lost.

At the end of 2009 the original project manager, Luis Martinez-Uribe, left the University of Oxford to take up a new role at the Instituto Juan March in Spain, being replaced by James A. J. Wilson. Fortunately, Luis agreed to continue working for EIDCSR on a one-day-a-week consultancy basis after 2009, which greatly assisted the continuity of the project, although there was inevitably a period of familiarization after the new project manager took over.

To add to the complexities of managing the project, the Bodleian Libraries representative leading the metadata management work left for a new position in July 2010 and staffing resources at the Libraries were too stretched for them to be able to invest the expertise in EIDCSR that had initially been hoped for.

The possibility of recruitment difficulties and losing staff part-way through the project had of course been identified in the initial project risk assessment and the mitigating actions proposed were followed in practice, but the disruption did result in the technical outputs not being as polished as initially hoped. Despite this, the project achieved its initial objectives and almost all of the planned deliverables were indeed delivered.

[10] See http://www.dcc.ac.uk/resources/tools-and-applications/data-asset-framework.

Work undertaken and milestones achieved in 2009

- April 2009: EIDCSR Project website, blog, and bookmarking site established (http://eidcsr.oucs.ox.ac.uk; http://eidcsr.blogspot.com; www.diigo.com/profile/eidcsr)
- July 2009: Response to the DCC Data Management Plan Content Checklist prepared and circulated, providing suggestions as to how the checklist could be improved (http://eidcsr.oucs.ox.ac.uk/docs/EIDCSR Response to the DCC DMP.pdf)
- July 2009: An appropriate evaluator was identified to conduct a formative evaluation of the project
- July-August 2009: A successful proposal for the Supporting Data Management Infrastructure for the Humanities (Sudamih) project in response to the JISC Research Data Programme, Data Management Infrastructure Call for Projects was prepared and submitted.
- August 2009: The initial draft of the Data Audit and Requirements Analysis Findings was produced and sections circulated to the researchers who contributed
- October 2009: First project workshop staged, entitled 'From Lab to Reuse' (http:eidcsr.oucs.ox.ac.uk/workshop.xml). Participants in the workshop had the opportunity to learn about, and contribute to discussions of, the different approaches to the ensuring the flow of data between laboratory and *in silico* experimentation. In particular, the workshop discussed: methods for the capture, storage and reuse of metadata in the laboratory; lifecycles integrating wet lab and *in silico* experimental data; the delivery and visualisation of large-scale data.
- November 2009: EIDCSR completes and submits its cost models report for the Keeping Research Data Safe 2 Project. This consists of an analysis of the relative costs of long-term data preservation and curation compared with the costs of data curation.
- November 2009: Research Services initiates its policy development work, seconding Paul Taylor from the University of Melbourne to draft a policy, drawing upon his experiences there.
- December 2009: A 'high-level' technical analysis document is produced, clarifying data workflows between the departments involved in the project and suggesting use cases.

Work undertaken and milestones achieved in 2010

- January 2010: Technical developer hired, enabling work to commence on translating technical analysis document into technical requirements
- March 2010: Research Services receives a report of Paul Taylor's secondment, also proposing a draft research data management policy for the University, which is refined and redrafted over the coming months.
- February-July 2010: Development of a new archiving client for the Computing Services's Hierarchical File Server, enabling the user-friendly archiving of very large quantities of data alongside metadata.
- March 2010: Second project workshop staged, focusing on the development of institutional policy and guidance for research data (http://eidcsr.oucs.ox.ac.uk/policy_workshop.xml). The experiences of the Universities of Oxford, Melbourne, Southampton, and Edinburgh were related, along with a funders' perspective from the BBSRC.
- March-April 2010: Evaluation of project progress undertaken by Angus Whyte of the Digital Curation Centre (DCC). Recommendation are made regarding how the project can better engage with researchers and improve awareness of what is meant by embedding curation practices.
- March-September 2010: A second software developer, Tahir Mansoori is employed to develop visualisation software that will help the 3D Heart Project researchers (and potentially others) view, share, and annotate very large 2D and 3D images.
- April 2010: Metadata workflow established (later enhanced)
- June - July 2010: Metadata schema developed to try to capture essential information regarding research datasets to assist rediscovery and reuse. This is then circulated to the Libraries and the researchers on the 3D Heart Project for feedback.
- July 2010: Project extension applied for and project timetable revised and updated
- August-November 2010: Research data management web portal created and made available to public (http://www.admin.ox.ac.uk/rdm/)

- September-October 2010: Scenarios document written and circulated to researchers, illustrating in a step-by-step manner how various data management tasks would be undertaken using the system being developed. Scenarios included setting up a new research project, adding metadata to a data set, archiving metadata, and using an interface to search for and retrieve datasets for re-use.
- October 2010: Draft Service Level Description produced to describe future service proposal and as a basis for service costing
- November 2010: Creation of online metadata editing interface, circulated to researchers for feedback
- November – December 2010: completion of software client for packaging and sending metadata to Libraries' 'Databank' system (built on the DAMS infrastructure)
- December 2010: Implementation of basic metadata browse and retrieval interface for Databank
- December 2010: Completion of cost/benefits analysis for future development and implementation of EIDCSR data preservation and rediscovery service

Outputs and Results

Understanding Researcher Requirements

Besides the initial project dissemination work, detailed in Appendix A, the first phase of the project involved a requirements-gathering exercise and an audit of the data assets of the 3D Heart Project and the processes by which they were derived. The 'Audit and Requirements Analysis Findings' report that resulted from this set out the context of the project and provided a stage-by-stage summary of each part in a chain of experimental procedures. Each heart begins by undergoing anatomical and diffusion tensor magnetic resonance Imaging (MRI) scans to produce 3D images. Then the heart is sliced and stained, and 2D histological images are taken with microscopes. After that the images are segmented before a 3D mesh is generated upon which *in silico* experiments can be conducted. Simulations may then be conducted on the heart models. A further use of the models is in the creation of a 'probabilistic 3D Heart Atlas'. Different research groups are involved in different stages of the process, although many of the requirements are similar. The requirements identified may be summarised as follows:

- Regular back-up of data with copies held securely off site
- Version control capabilities to track changes to data
- The ability to restrict access to in-development data and share the final outputs
- Fast access to the (very large) histology data, including the ability to visualize parts of it
- A means to make data available to collaborators
- The ability to search for stored data
- The ability to access information about experiments and simulations remotely
- And ideally the ability to publish articles based on data without having to go back to log-books, notes, and files to find information allowing reproducibility

These requirements informed the subsequent phases of the project.

It is worth bearing in mind that whilst these requirements came from interviews with the researchers involved in the 3D Heart Project, not all of the researchers could immediately see the benefits that might come from data rediscovery and re-use – preservation was the primary concern for many. If data is to be re-used, it needs to be documented; and if data is to be properly documented by the researchers that create it, they need understand why they are being asked to document it. This is an area where the need for policy and training becomes apparent.

Whilst the immediate goal of EIDCSR was to assist the research groups who were giving up their time to work with the project, it was always the intention that as an infrastructure project EIDCSR would keep one eye on the 'generalisability' of both the requirements it sought to meet and the components it developed. From the preliminary scoping work, we felt that the requirements of the researchers involved in the 3D Heart Project were likely to be relatively typical, although we decided that resources might in the first instance most profitably be directed towards secure preservation, access

controls and availability, and implementing a system of documentation that would enable searching and access to metadata describing the experiments.

Development of institutional Policy

In 2005 the University of Melbourne published a draft institutional 'Policy on the Management of Research Data and Records', which was intended 'to assist departments and individual researchers to fulfil their responsibilities with respect to the storage and retention of data and records associated with, and arising from, their research activities'.[11] Given that EIDCSR intended to develop a similar institutional policy for Oxford, the Project seconded a senior staff member from Melbourne's Research Office to interview key stakeholders in Oxford and to draft a suitable policy. Key recommendations arising from this work included:

- That the University of Oxford approves an institutional policy on the management of research data and records, and recommends that departments (and maybe Divisions) develop local, contextualised discipline-specific versions of this policy and procedures for this to be implemented.
- That the University embed the institutional policy requirements in conditions of award for internal grant/funding schemes
- That the University develops a 'research data management portal', which would be a one-stop shop for information about research data management requirements and services.
- That the University considers resourcing a Data Management Advisory Service.
- And, that a Data Managers Forum is established to enable sharing of experiences, best practice, problems and their solution across the University.

Research Services re-drafted and edited the Melbourne recommendations after additional consultation during 2010, and a concise three-page 'Draft Policy on the Management of Research Data and Records' will be submitted to the University's Research Information Management Sub-committee (RIMSC) at its meeting in February 2011, before this is considered by the University's Research Committee. Although the policy is not yet ready for publication, its broad outline involves: defining the purpose of the policy and its scope; describing the principles upon which the policy is based; indicating the particular data management responsibilities of the University as a corporate body, the responsibilities of the departments within it, and the responsibilities of individual researchers; and mentioning how the policy operates in the context of existing policies at the University. It is clear from the policy that data management is seen as a process that needs to involve multiple agents at various levels of the institution: "Research data and records management is a shared responsibility, and researchers, Departments, central administrative units and service providers need to work in partnership to implement good practice".

Data Management and Curation Advice & Guidance

Although not part of the initial EIDCSR project plan, one of the recommendations to come out of the Institutional Policy Development work package was that the University should develop some sort of central data management Web portal to provide researchers with advice and guidance (e.g. about researcher funders' data management requirement and policies, how to produce a data management plan etc.), and to signpost them to relevant support services and training. It was felt that collating and presenting important information about data management in one place would help to improve practice in data management and might also help the University to identify services where further development is required Research Services therefore used some of their EIDCSR policy development budget allocation to lead on developing the portal, with additional content contributed by project staff at the Computing Services.

The University of Oxford Research Data Management Web Portal (http://www.admin.ox.ac.uk/rdm/) takes a 'whole life-cycle' approach to data management and curation, with a particular emphasis on the planning and preparation stage (befitting the expertise of Research Services). The home page

[11] http://www.unimelb.edu.au/records/research.html

presents the various stages of the life-cycle via a wheel, which users can click to drill down to greater detail (see appendix B).

Besides the portal, EIDCSR has also developed an introductory data management leaflet for researchers, which refers them to services both within and external to Oxford that can help them with various aspects of data management. We hope that other Universities may re-use it and adapt it for their own institution. We have been handing out the leaflet at induction events for new researchers, and it is also available from the project website (http://eidcsr.oucs.ox.ac.uk/docs/Research Data Management Leaflet v 1 25.pdf).

It is at present too early to judge the impact of the web portal, which was made live in November 2010, although it does provide the central reference point for researchers that would almost certainly need to be provided by any University wishing to establish a researcher-focussed data management infrastructure.

Development of Secure Storage and Metadata Management Service

Essential to the EIDCSR Project was the development of a workflow that could embed data curation practices within research over the course of the data life-cycle. This needed to be done in such a way as to address the key requirements of the researchers on the 3D Heart Project whilst ensuring that the workflows described and the interventions made were general enough to be more broadly applicable across research groups at the University. It was planned from the outset that the Hierarchical File Server (HFS) at the Computing Services would be used to provide very secure long-term data storage, and anticipated that the Libraries' Fedora Digital Asset Management (DAM) system would provide appropriate metadata management infrastructure. The IBM HFS long-term file store was already a mature service that had been in operation for several years by the start of EIDCSR. The DAM system, however, was still very much in development, and remains so at the time of writing this report. The service name 'Databank' has been given to the data storage and management functionality of the Oxford DAM system, and this shall be used hereafter.

The data curation workflow upon which the EIDCSR project settled may be mapped to the required elements of data management infrastructure diagram provided under the 'methodology' section of this report (albeit with the 'project set-up' phase broken out from the project planning phase):

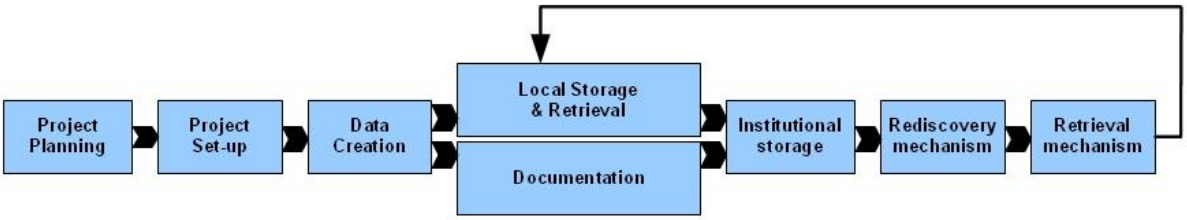


Figure 1. Generic steps of research data curation

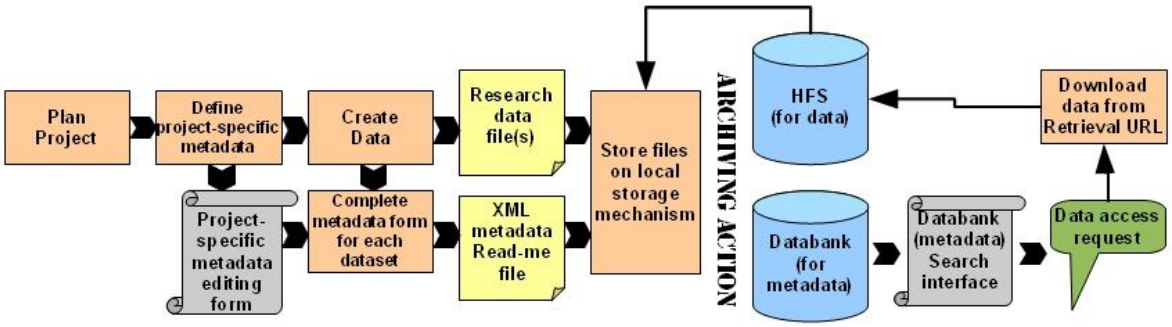


Figure 2. EIDCSR workflow interventions

The orange-shaded boxes in figure 2 indicate researcher actions; the yellow icons indicate computer files; the blue cylinders indicate the elements of institutional infrastructure where data is stored; and the grey scrolls represent online interfaces for metadata description and searching.

Each research project is expected to begin with a data planning phase, although EIDCSR did not seek to intervene in this stage directly (the 3D Heart Project had already been funded in any case). The EIDCSR workflow assumes that during the set-up phase of a project the PI will add to a basic core set of metadata fields any project-specific fields that will be useful to those wishing to describe the data being created or find it again in future. For instance, in the 3D Heart Project this would include information about the species, sex, and weight of the specimen from which the heart was acquired. Once the project is fully underway and the research process is generating data, the researchers are expected to complete a copy of the metadata form defined in the previous phase for each dataset they produce, thereby providing documentation. The metadata form outputs a short piece of XML code which the researcher simply saves to the local directory where the data described resides. When the data archiving client is used to archive the data (which may be done regularly during the course of a project), the person doing the archiving simultaneously activates the EIDCSR metadata archiving client: the data files are associated with the metadata, and whilst the former are sent to the HFS system, the latter are sent to Databank. The Databank search interface can then be used to locate relevant data resources, view the metadata describing each dataset (including the process by which it was generated), and the researcher (or in due course a member of the public if the data is not marked as restricted) may place a request to access the data. If this request is approved by the designated data curator, the requestor will be sent a URL from which they may download the requested data. The data curator is set by default to be the PI of the research project, but this role may be delegated to another researcher or somebody with a more specific data curator position.

In order to facilitate the workflow described above, EIDCSR has produced: a core research metadata schema; a metadata addition/editing webform for the 3D Heart Project; a new version of the IBM archiving client for archiving data; a lightweight metadata-archiving client; and a basic Databank search/retrieval interface. Due to resource and time restrictions these tools are presently only at proof-of-concept stage. They will be used by the 3D Heart Project, but lack the polish to be rolled out more broadly as yet. The University will consider funding to trial the tools more broadly and bring them up to production standard in due course.[12]

We have adopted the approach of storing data and metadata separately in order to make the best use of existing infrastructure and keep costs down. Other universities undertaking similar exercises may wish to consider keeping the data and metadata together to simplify processes. As the IBM HFS system at Oxford only provides one fairly limited field for metadata for each file or directory we opted to use this to store not the metadata itself, but a unique identifier to link the data with the metadata.

Metadata

EIDCSR worked with the Bodleian Libraries to identify existing metadata standards which could be used to document research data. Whilst there are a number of discipline-specific standards that might have been used, finding a basic metadata schema that could be applied equally to almost any research data proved difficult. EIDCSR therefore opted to create its own core set of metadata fields whilst allowing individual research groups to define their own fields that would be relevant to their particular interests. This resulted in the development of an extensible set of XML fields, and a tool to enable a project PI to customize the metadata entry form to include additional fields at their discretion.

The Libraries' Databank system, whilst extremely flexible, required that any metadata sent to it had minimum of four fields of information: a unique identifier; a named data creator; rights information; and a list of files comprising the dataset. Beyond that, we considered elements from the Dublin Core terms set,[13] and (once a draft was available) the proposed DataCite schema.[14] We shared our list of

[12] See appendix C for the estimated costs of investment required to bring the proposed service to production standard.
[13] See the Dublin Core Metadata Intiative web pages: http://dublincore.org/documents/dcmi-terms/.
[14] See http://datacite.org/schema/DataCite-MetadataKernel_v2.0.pdf.

proposed core fields with the researchers on the 3D Heart Project to get their feedback and input, and settled on the following draft list of core fields (with explanations):

1. **[unique ID#]**
2. **Data Creator(s)** *(to include the name and affiliation of each person involved in creating the data being described)*
3. **Rights information** *(guidelines as to how the data can be used, including the copyright holder)*
4. **File(s) comprising dataset**
5. **Title** *(of dataset)*
6. Project Name *(The project under which the data was generated)*
7. Funding Agency *(The agency that funded the project, if applicable)*
8. Grant Number *(The grant number of the project, if applicable)*
9. **Description** *(free text description of what the data actually represents)*
10. **Process** *(free text description of how the data was generated – including experimental details such as hardware/software used and input parameters, location of facilities used, etc. In practice, where projects are applying consistent parameters they may wish to define these as additional project-specific metadata fields, in order to improve searchability)*
11. Relationships *(includes type of relationship (using the Dublin Core relationships options) and a reference to some other data/output, e.g. provenance data if the dataset was based on earlier data, for instance each 3D Heart Project MESH dataset would refer to the Segmentation dataset it was based on, which would in turn refer back to the Histology and MRI datasets that the Segmentation dataset were based on. The field could also reference published articles or presentations that have referred to the dataset)*
12. Dates of data capture – Start and End *(this can be left open-ended if data is still being added to the dataset)*
13. Keywords
14. Language *(defaults to English)*

Curation metadata:

15. Access restrictions
16. Public release date *(if applicable)*
17. Data destruction/review date
18. Data Curator
19. Data Curator contact details
20. [Date of data deposit]
21. [file format(s)]

Fields in square brackets are those that could be generated automatically by the system. Some other fields, whilst editable, will in practice default to values given at the project set-up stage (Project Name, Funding Agency, and Grant no. are all in this category). Those fields listed in bold are required; the others are optional.

It is worth drawing attention to the 'Process' field, which is in addition to a standard Dublin Core 'Description' field. Whilst the Description is intended for a description of what the data represents, the Process is intended to record how the data was derived. Although in practice few researchers are likely to write such a full account of the processes involved in the generation of the data to render it properly reproducible, the inclusion of this field is intended to go some way towards addressing that requirement.

The challenge when producing the core schema was to gather as much useful data about a given dataset as possible, without overburdening the researcher who would have the job of recording the information. Given that there is no means for the University to effectively enforce data documentation, the demands placed upon the researcher needed to be realistic. In order to assess this, EIDCSR put together a preliminary metadata editing form, populated with the core fields and a few project-specific

ones to the researchers working with the project, and asked them to answer a number of short questions about the form. The responses indicated that most researchers felt that the process would be manageable in practice, although small improvements in design (particularly the use of drop-down menus for responses and suchlike) and automatically populating some fields (which was planned, but not implemented in the test form) could make a significant difference to researcher's willingness to complete it. Most researchers, when asked to complete the form for a real dataset they had created, reported that it took between ten and fifteen minutes, which would be brief enough to incorporate into their standard research practices.

The metadata schema and editing interface would need to be tested with researchers from other disciplines before being rolled out more widely.

Visualisation tools

Whilst the EIDCSR Project was ostensibly about the preservation of research data, perhaps surprisingly the requirements specified by the research groups emphasised the importance of access to the data (for which preservation was one means). Nor was access to data a longer-term future possibility but rather a requirement within the existing collaborative research activiites (where the collaboration was distributed across continents). Therefore, although not included in the generalised research data workflow given above, nor indeed in the original EIDCSR Project Plan, the 3D Heart Project required a means by which the scientists could rapidly visualize and share details from the very-high resolution (and therefore bandwidth-intensive) heart images they were creating. To this end, EIDCSR recruited one of the researchers from the Oxford Computing Laboratory who was already working with the project to further develop some visualisation software that he had been working on during his doctoral research.

This 'Workbench' software consisted of a web-based interface for viewing and zooming in to the huge two-dimensional histology and three-dimensional MRI images that the Department of Physiology, Anatomy, and Genetics and the Department of Cardiovascular Medicine were generating. The image files themselves were stored on a server and sub-divided into many small parts at progressively higher resolutions, enabling fast, smooth, and seamless visualisation. The software enabled the images and annotations of the images to be accessed (with fine-grained access controls) by collaborators in New Zealand (amongst others), and also provided thresholding tools so that particular features of the images could be brought out and emphasised.

The software was demonstrated at Oxford University Computing Services where it was felt that it may have applications beyond the sciences – to scholars involved in examining images of manuscripts or archaeological artefacts for instance. It is entirely open source, so it can be taken and adapted by other institutions. At the time of writing we are adding documentation, seeking an appropriate place to host the code, and considering how to draw it to the attention of a wider development community (including a consideration of its relevance to the related Sudamih project).

Business and Cost Models

The business and cost modelling activities of the EIDCSR Project were divided into two distinct phases. During 2009, EIDCSR worked with the Keeping Research Data Safe 2 (KRDS2) Project to provide a case study, contributing detailed cost information about the creation, local management, and curation of the data produced by the 3D Heart Project. The Oxford case study for KRDS2 nicely complemented the other contributions, mostly from data centres, where the management of data was understood in terms of an OAIS model. Oxford represented a case where the management and curation of research data is undertaken as an institutional collaborative effort involving researchers and service providers from conception of the research project.

The results of the exercise showed that the cost of creating the actual research datasets, including staff time as well as the use and acquisition of lab equipment, was proportionally high, representing 73% of the total cost. The costs of the admittedly limited local data management activities undertaken by the researchers on the other hand were modest, representing only 1% of the total. The curatorial activities undertaken as part of the EIDCSR project constituted 24% of the total, but much of this

represented start-up costs which would not need to be factored into the equation were data curation infrastructure already in place. 2% of the costs stemmed from the secure storage of the very large datasets (c. 5TB) for five years on the long-term file store. Although these proportions are specific to the project, they are likely to be broadly indicative, and one would certainly expect to achieve significant economies of scale once the required infrastructure is in place.

In the latter part of 2010, EIDCSR began to look at future service costs if the 'Secure Storage and Metadata Management Service' being developed by the project were to be rolled out across the University in a sustainable manner. One of the advantages of building the new service around existing infrastructure components was that prices for several components (on a cost-recovery basis) were already set. Obviously, were the new service to be introduced and significantly scaled up, some of these prices might need to be altered to reflect either increasing support costs or economies of scale, necessitating the recalculation of component costs, but for basic costing purposes we were not starting from a position of ignorance.

A Service Level Description was drafted to describe the envisaged 'Secure Storage and Metadata Management Service', what it does, who can use it, and the clients' responsibilities. The SLD was then used as a framework for cost calculations. As these calculations mostly drew on existing HFS and Databank service prices, which are not charged in a consistent way, the resulting model is not entirely straightforward. The costs of establishing and maintaining the future service were characterised into five groups: additional capital development costs required to complete the service infrastructure; fixed costs per year of maintaining the service infrastructure; variable costs associated with each research project; variable costs per TB of data stored; and variable costs per dataset documented. A full breakdown of the estimated costs of establishing and running the proposed service is provided in Appendix C.[15]

Whilst the costs of implementing the Secure Storage and Metadata Management Service are relatively straightforward to estimate, it is not so easy to quantify the benefits. This is especially the case given that the re-creation value of data is likely to vary considerably from research project to research project, and does not correlate to the raw size of the data preserved – it would take far longer and be more expensive, for example, to recreate the relatively small database of the Roman Economy Project (which the Sudamih project is working with), than to recreate the large-scale histological data for a rat heart for instance (although the new rat heart would not of course be exactly the same as the previous one).

The KRDS2 final report recommends using a benefits taxonomy to identify benefits arising from research data curation.[16] This considers direct and indirect benefits as one dimension, near-term and long-term benefits as another, and private and public benefits as a third.

As far as the EIDCSR Project itself is concerned, the direct benefits are felt by the research groups that participated in the project, who have as a result fulfilled the BBSRC data preservation mandate and improved productivity due to: datasets being easier to locate; the development of a visualization interface that enables them to access and analyse the large image datasets; and the consolidation of data back-up. These benefits may be best understood as costs avoided. Some of the data produced by researchers in the 3D Heart Project would be expensive to recreate (even if the REP database would be more expensive still), but when such data is secure and re-discoverable then no re-creation is necessary. Furthermore, by enabling the researchers themselves to manage and describe their data at the time of its creation, there is no need to curate the data long after it was originally created, which, as the KRDS2 report has shown, tends to be much more expensive.

The near-term benefits are the tangible and practical benefits to the researchers participating in EIDCSR, which include secure data storage, search capabilities, improved data access times, and easier short-term reuse of well-managed data. The long-term benefits are arguably more significant as they allow the data to be re-used by future research projects staffed by different researchers. Without long-term storage and searchable documentation, such re-use depends largely on personal

[15] The University of Oxford has been working with the JISC to develop a process to measure costs of infrastructure services, which has loosely informed this costing. The 'Toolkit for Costing IT Services' has not yet been published.

[16] Beagrie, N., Lavoie, B., Woollard, M., 'Keeping Research Data Safe 2', 2010, pp. 53-63. http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf.

contacts and the ability of the original researchers to actually find data that is frequently simply recorded on DVDs and placed in office drawers.

Beyond the private benefits to the researchers who created the data, the BBSRC benefits as the value of the data created as a result of its investment of public money can be maximised, service providers at Oxford benefit from a better understanding of the needs of researchers and the applicability of the tools developed, and institutions beyond Oxford hopefully benefit from the experiences of the EIDCSR project and lessons learnt.

Dissemination Activities & Community Engagement

A website was established in order to communicate the EIDCSR Project's goals and host project outputs, whilst a project blog was set up to provide a mechanism for alerting the data curation community to new developments and issues of interest. In addition to these, a bookmarking service was set up to provide links to related websites.

Two workshops were staged, the first, 'From Lab to Reuse', was used as a chance for the project to learn more about data flows in laboratory environments. The second workshop, 'Institutional Policy and Guidance for Research Data' provided a forum in which the University of Oxford (along with colleagues at the University of Melbourne) could inform other universities about the approach we were taking, whilst also learning more about what other institutions were doing in regards to developing policy.

Over the course of the project EIDCSR staff gave a number of presentations and published several articles explaining the work being undertaken, the approach informing it, and the findings the projects was uncovering. A full list of dissemination activities is provided in Appendix A.

Outcomes

The EIDCSR Project achieved its objectives in that there is now sufficient core infrastructure in place to enable the research groups involved in the 3D Heart Project to document their data, store it securely off-site, search the documentation, and retrieve data from the long-term file storage. In some respects the project achieved more than it had initially set out to, developing data visualisation software and a researcher-facing data management web portal. Due to the delay in recruiting appropriately-qualified technical staff, however, the project did not have the opportunity to apply the polish required to the tools it developed in order to render them 'production ready'.

Project Objectives met:

- EIDCSR produced a system, with supporting tools, for the researchers involved in the 3D Heart Project to utilize the HFS and Databank services for data preservation. The generalised research data workflow and curation tools should prove applicable and flexible enough to enable research groups from other disciplines to similarly preserve and document their research outputs.
- Sheer curatorial practices have been introduced to existing research workflows in order to add value to everyday scholarly work. The metadata entry forms and combination of data and metadata archiving tools are the principal examples of this.
- Work on developing institutional policy and guidelines has resulted in a draft policy document and outline implementation plan which will soon be considered by the University's Research Information Management Committee and then the Research Committee; a service level description has been drafted to illustrate what a 'Secure Storage and Metadata Management Service' would offer in practice, and a model developed setting out the costs of such a service.
- The project has considered the roles and responsibilities of service providers in Oxford to support their researchers by developing workflows and the stages of the data lifecycle in which particular services can intervene most profitably.

The specific deliverable outlined in the 'aims and objectives' section of this report have all been completed, with the exception of the report on the strategic provision of filestore and access management infrastructure (planned to be undertaken with IBM). The software behind the metadata entry forms and Databank searching interface, originally intended to be released under an open source licence by the end of the project, is not yet polished to a degree that would justify its open distribution without further development, although the visualisation software developed by EIDCSR should be ready for release under an open source licence during early 2011.

The work undertaken during the EIDCSR project should, with a little further development, have a huge impact on the research community within the University of Oxford and, by extension, other UK universities. For the first time, researchers in fields that are not already served by national or international data centres will have somewhere to securely deposit their data outputs, tools for describing, finding, and re-using them, and the ability to (manually) link their data to research outputs derived from it. By avoiding the risk of data loss and facilitating data re-use, the value of research data will be significantly enhanced. The approach to developing data management infrastructure at Oxford can hopefully inform decisions elsewhere and provide an example for others to learn from.

Conclusions & Implications

The EIDCSR Project provides an exemplar for other institutions wishing to develop their own data management infrastructures. The project illustrates how existing services and support groups can be adapted and enhanced, at relatively low cost, to work together to provide a basic but joined-up research data management solution that can help data through the life-cycle from conception through to long-term preservation and re-use.

The University of Oxford has taken the approach that it is easier to construct its data management infrastructure around aspects of infrastructure, both technical and service, that are already in place. Other universities are likely to share several aspects of existing infrastructure, but there will inevitably be some differences between the situation here and elsewhere. As such, it may not be viable (or even desirable) for other institutions to try to repeat the work undertaken by EIDCSR exactly. We have therefore tried to draw out some general lessons from our experiences that we think will be generally applicable to others, and which hopefully will help smooth the process:

- It is important to work with researchers to ensure that any infrastructure or workflows designed for them will in practice be implementable. Maintaining 'buy-in' to such a project can however be a challenge. Clearly identify who you will be working with and ensure they are involved in all significant decisions.
- Don't assume that your researchers think a great deal about the potential re-use of their data. Most researchers are primarily interested in the publication of their results, as this is how they gain career-enhancing recognition and reward. You may need to present a clear case to them regarding the value of their data and why it is worth managing properly. It is safer to appeal to self-interest (i.e. saving time and avoiding risks of data loss) than to altruism.
- Try to work as much as possible with existing research work-flows. Your researchers are unlikely to be impressed if ordered to fundamentally change the way they conduct their research. Ensure that interventions in workflows are acceptable and not over-burdensome.
- Making use of and enhancing existing institutional technical infrastructure and services may help keep the costs of supporting research data management down.
- Where metadata is concerned there may be a trade-off between comprehensiveness and usability. It is more important that some documentation takes place rather than none.
- There is a balancing act between meeting the immediate needs of the researchers you are working with whilst ensuring the project outputs are generalisable. To much emphasis on one or the other will result either in alienating your most important collaborators or producing an infrastructure that is not well suited to the overall needs of your institution
- Communication between service providers within an institution is also essential. Given that the data management lifecycle touches upon many parts of a university's support services, they each need to be aware of any data management infrastructure development activities, and if possible given early and clear responsibilities in such developments, to ensure engagement and buy-in.

- Closed APIs for existing software infrastructure can slow technical development. Know what it is you are dealing when adapting infrastructure and estimate development time accordingly.
- Ensure that any technical modifications to existing infrastructure can be supported by the group with responsibility for that infrastructure.
- Build recruitment time into project plans and remember that it's easier to recruit technical staff when you offer them a full-time position(!)
- When working with external partners (such as IBM), make sure it's clear what their role will be from the outset and who precisely you need to work with
- Universities often take decisions slowly. New policies can take time to be approved, and some aspects of infrastructure implementation may need to be delayed until policy is settled.
- When developing institutional data management policies take into account the need to raise awareness amongst researchers and point them to the right resources and support within the institution and beyond.
- Quantifying the benefits of research data curation is a complicated business. Current work being undertaken by the JISC Managing Research Data Programme to assess costs and benefits is likely to prove helpful when developing business cases.

Whilst the EIDCSR project has made strides towards the implementation of an institutional Secure Storage and Metadata Management Service for researchers, more work needs to be undertaken to ensure that the model we envisage is appropriate across academic disciplines, and it is possible that refinements will need to be made even though the customisable nature of the metadata collected provides an in-built degree of flexibility. That said, the broad approach taken by EIDCSR, and the core metadata schema it has settled upon, are likely to be generally applicable to other Universities intending to develop their own cross-disciplinary research data management infrastructures and we hope our work can inform future projects.

Appendix A: Dissemination Activities and Outputs

Websites:

- EIDCSR Project website (http://eidcsr.oucs.ox.ac.uk)
- EIDCSR Project blog (http://eidcsr.blogspot.com)
- EIDCSR Project bookmarking site (www.diigo.com/profile/eidcsr)
- Dodd, Sian; Dally, Kathryn (eds.), 'University of Oxford Research Data Management Portal', http://www.admin.ox.ac.uk/drm, November 2010.

Workshops

- EIDCSR first data curation workshop: 'From Lab to Reuse', 14 October 2009. http://eidcsr.oucs.ox.ac.uk/workshop.xml
- EIDCSR second data curation workshop: ''Institutional Policy and Guidance for Research Data', 29 March 2010. http://eidcsr.oucs.ox.ac.uk/policy_workshop.xml

Software:

- Akram, Asif, 'EIDCSR IBM Tivoli Storage Manager-compatable archiving client', 2010
- Akram, Asif, 'EIDCSR/SUDAMIH form-builder software', 2010
- Akram. Asif, 'EIDCSR Metadata archiving client', 2010
- Mansoori, Tahir, 'Workbench 2D and 3D Visualisation Software', 2010

Publications:

- Martinez-Uribe, L., Macdonald S. 'User engagement in research data curation', in *Research and Advanced Technology for Digital Libraries, 13th European Conference, ECDL 2009 Proceedings*, Lecture Notes in Computer Science 5714, October 2009. http://www.springerlink.com/content/7mnq13x34717p483/
- Martinez-Uribe, L., 'Repositoris de dades a Oxford: serveis institucionals federats', Teraflop 105 December 2009. http://www.cesca.es/promocio/teraflop/tera105.pdf
- Martinez-Uribe, L. Macdonald, S. 'Collaboration to data curation: harnessing institutional expertise' New Review of Academic Librarianship Vol. 16 (S1), October 2010.
- Wilson, James A. J., Fraser, Michael A., Martinez-Uribe, L., Jeffreys, P., Patrick, M., Akram, A., Mansoori, T., 'Developing Infrastructure for Research Data Management at the University of Oxford' Ariadne 65, October 2010. http://www.ariadne.ac.uk/issue65/wilson-et-al/
- Wilson, James A. J., Martinez-Uribe, L., Fraser, Michael A., Jeffreys, P., 'An Institutional Approach to Developing Research Data Management Infrastructure', International Journal of Digital Curation (forthcoming).

Presentations:

- Martinez-Uribe, L., 'Scoping and Developing Institutional Data Services: the Data Libraries of 2020', IASSIST 2009, Tampere, Finland. 29th May 2009. (http://www.fsd.uta.fi/iassist2009/presentations/F1_Martinez.pdf)
- Martinez-Uribe, L., 'The application of the data audit framework (DAF) within DataShare at Oxford', DAF review, Edinburgh. 10th June 2009. (http://www.slideshare.net/luismartinezuribe/the-application-of-the-data-audit-framework-daf-within-datashare-at-oxford)
- Jeffreys, P. W., 'Data imperative workshop: welcome and introduction', Data Imperative Event, Oxford. 3rd June 2009. (http://www.ict.ox.ac.uk/odit/projects/digitalrepository/docs/DataImp030609-PaulJeffreys.pdf)

- Martinez-Uribe, L., 'Research data management at the University of Oxford'. Data Imperative Event, Oxford. 3rd June 2009. (http://www.ict.ox.ac.uk/odit/projects/digitalrepository/docs/DataImp030609-LuisMartinez.pdf)
- Martinez-Uribe, L. and Macdonald. S. 'User engagement in research data curation', European Conference on Digital Libraries (ECDL) 2009, Corfu. 30th September 2009.
- Martinez-Uribe, L. 'Data Repositories in Oxford: Institutional Federation of Services', Jornada de Supercomputacion, 2nd December 2009. (http://eidcsr.oucs.ox.ac.uk/docs/JOCS%20-%20LuisMartinez.pdf)
- Martinez-Uribe, L. EIDCSR Poster at the DCC conference: 'Moving to Multi-Scale Science: Managing Complexity and Diversity', London, 3rd December 2009.
- Martinez-Uribe, L. & Rumsey, Sally, 'Digital data preservation and the role of the libraries', All Hands Meeting 2009, Oxford, 7th December.
- Jeffries, Paul, 'Research data management: can we do it?', Disseminating Oxford's Research in the Era of Electronic Communication seminar, Oxford, 4th March 2010.
- Wilson, James A. J., 'EIDCSR Project Update', Preservation Exemplar Projects Workshop, London, 14th June 2010.
- Wilson, James A. J., 'Everything You Wanted to Know about OUCS Research Data Management Projects', Oxford, 7th July 2010.
- Martinez-Uribe, L. 'Implementing data repository services: issues and lessons learned from case studies', Databases in Quantum Chemistry: Validation of methods and software, and repositories of reference computational results, Zaragoza, 23rd September 2010. (http://neptuno.unizar.es/events/qcdatabases2010/files/Luis_Martinez-Uribe.pdf)
- Wilson, James A. J., Briefing about data curation education activities at Oxford, Vienna, 22nd September 2010.
- Wilson, James A. J., Fraser, Michael A., Martinez-Uribe, L., Jeffreys, P., 'An Institutional Approach to Developing Research Data Management Infrastructure', 6th International Digital Curation Conference, Chicago, 7th December 2010. (http://eidcsr.oucs.ox.ac.uk/docs/OxfordInfrastructure_IDCC2010.pdf)
- Wilson, James A. J., 'Embedding Institutional Data Curation Services in Research (EIDCSR)', Oxford, 14th January 2011. http://eidcsr.oucs.ox.ac.uk/docs/EIDCSRforOUCS_jan2011.pdf

Reports:

- Martinez-Uribe, L., 'Response to DCC Data Management Content Checklist', 2009. (http://eidcsr.oucs.ox.ac.uk/docs/EIDCSR%20Response%20to%20the%20DCC%20DMP.pdf)
- Martinez-Uribe, L., 'Data curation: from lab to reuse. Workshop Report', October 2009. (http://eidcsr.oucs.ox.ac.uk/docs/EIDCSRWorkshop14-10-09%20-%20Report.pdf)
- Martinez-Uribe, L., 'EIDCSR Audit and requirements analysis report', December, 2009. (http://eidcsr.oucs.ox.ac.uk/docs/EIDCSR_AnalysisFindings_v3.1.pdf)
- Martinez-Uribe, L., Wilson, J. A. J., 'Institutional Policy and Guidance for Research Data - Workshop Report', April 2010. (http://eidcsr.oucs.ox.ac.uk/docs/EIDCSR_Workshop_29_Mar_10_Report.pdf)
- Martinez-Uribe, L., [Contribution to the] 'Keeping Research Data Safe 2 project final report', published May 2010. (http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf)

Support Materials:

- Dodd, S., Dally, K., Patrick, M., Wilson, James A. J., 'Managing Your Research Data at the University of Oxford' [leaflet], September 2010. (http://eidcsr.oucs.ox.ac.uk/docs/Research data leaflet.pdf)

Appendix B: The University of Oxford Research Data Management Web Portal

The following image shows a detail from the Research Data Management Portal developed by the EIDCSR project and Research Services. Each segment of the data management life-cycle 'wheel' directs the researcher to advice and links to further information.

Appendix C: Secure Storage and Metadata Management Service Cost Model

The following table estimates the additional capital expenditure required to bring the infrastructure developed during the EIDCSR project to a production service:

Additional pre-service development costs		*[All costs are calculated as FEC]*
Integration of archiving clients	£18,645	Grade 8, 2 months
Redesign of metadata forms & adaptation for hosting on OUCS Host service	£18,645	Grade 8, 2 months
System for editing metadata directly in Databank (primarily linking publications)	£18,645	Grade 8, 2 months
Development of user guide	£8,428	Grade 7, 1 month
Integration of visualisation tools with metadata	£27,967	Grade 8, 3 months
Trialling system with other research groups	£25,285	Grade 7, 3 months
Databank account	£5,000	
Hardware	£12,000	
TOTAL	*£134,685*	

The following table indicates the estimated annual fixed costs of running the 'Secure Storage and Metadata Management Service' and the associated variable costs for running the service at the level described by the draft service level description. The fixed costs are estimated on a per annum basis and include the ongoing research and development costs required to keep the various aspects of the infrastructure up to date, taking into account the anticipated changing expectations of users.

Estimated fixed costs of service per year		
Maintenance of Data Management Web portal	£4,862	Grade 7, 2.5 weeks
Promotion of policy	£2,150	Grade 8, 1 week
Maintenance of HFS/Metadata archiving clients	£8,605	Grade 8, 4 weeks
Maintenance of visualisation client	£4,303	Grade 8, 2 weeks
General systems administration & reporting	£3,890	Grade 7, 2 weeks
Hardware maintenance and support	£2,500	
TOTAL fixed costs of service per annum	*£26,310*	
Variable costs per project		
Project inception help and guidance	c. £100	Assuming 4 hours of support
User support for metadata entry	c. £100	Assuming 4 hours of support
Variable costs per TB of storage		
Storage in HFS	£842	per year per TB beyond first TB (price to externally funded projects)
Variable costs per dataset		
Databank metadata storage costs	Nominal	Databank service charges £5,000 for up to 1TB of storage in perpetuity. An average metadata file is likely to be less than 2K in size
Researcher time and effort	£13	assumes that 15 minutes of Grade 8 staff time (FEC) is spent creating each metadata file
DOI costs	c. £1	See the DOI price lists at mEDRA, for an example: http://www.medra.org/en/terms.htm

Were the 3D Heart Project to be paying for the (proposed) University of Oxford Secure Storage and Metadata Management Service, rather than working with the EIDCSR Project in order to design it, it would incur the following costs according to the model above:

3D Heart Project estimated data curation costs using proposed service cost model
5TB data stored in HFS for 5 years – £16,840 (n.b. first TB is free)
Assumed costs of central support for project – £200
Approximate number of hearts documented – 10
Approximate number of 'datasets' per heart – 8
Total costs of datasets – £1,120

TOTAL - £18,160

In contrast, the data curation needs of the Roman Economy Project (which the Sudamih data management infrastructure project is working with) would cost significantly less. Although the REP is a major research project, it has only one database requiring long-term preservation. The database is complex, but relatively small. As it is less than 1TB in size, it may be archived to the HFS without charge under the current pricing model.

Roman Economy Project estimated data curation costs using proposed service cost model
21MB data stored in HFS for 5 years - £0 (less than 1 TB)
Assumed costs of central support for project - £200
Number of datasets documented – 1
Total costs of datasets - £14
TOTAL - £214

As one can see from the above examples, the current pricing model effectively subsidizes projects with small data requirements at the expense of those with large data. It should be emphasised that the current pricing models used for the HFS do not reflect the true costs of the service. Long-term tape storage is relatively cheap; it is the staff costs involved in setting up HFS project accounts that in practice cost the most, so in reality the current terabyte price of £842 better reflects the actual cost to the institution of setting up an HFS account. As a result of the differences between price and cost the introduction of a new service such as that proposed by EIDCSR may in practice require a review of the current infrastructure pricing model.

The price model for the Databank service involves the initial up-front payment of £5,000 for up to 1TB of data. There are no annual fees beyond this initial charge. Given that the Databank will only be used to store metadata files, which are very small, rather than the underlying data the metadata describes (which will be stored in the HFS), a 1TB space allocation is anticipated to serve the 'Secure Storage and Metadata Management Service', acting as a broker for individual research projects, for a long time.

Of course, the fixed costs of maintaining the basic service (£23,810 per year) would need to be invested regardless of how many research projects actually used the service, so there would be moderate economies of scale to be gained by increasing usage of the proposed service.